Tiheysfunktion estimointi Bayes-SiZer -menetelmällä

Janne Koivunen
pro gradu -tutkielma
Matematiikan ja tilastotieteen laitos
Helsingin yliopisto
Huhtikuu 2004

Sisältö

1 Johdanto

Todennäköisyyslaskennassa ja tilastotieteessä satunnaismuuttujan tiheysfunktiolla on hyvin keskeinen rooli. Tiheysfunktion avulla on mahdollista ratkaista satunnaismuuttujaa koskevien tapahtumien todennäköisyydet ja näin saada selville satunnaismuuttujan todennäköisyysjakauma sekä siihen liittyvät tunnusluvut, kuten odotusarvo, varianssi ja vinous. Jo pelkät tiheysfunktion eksploratiiviset tarkastelut, kuten moodien lukumäärän ja sijaintien selvittäminen tiheysfunktion kuvallisesta esityksestä, antavat usein arvokasta tietoa satunnaismuuttujaan kvantifioidusta ilmiöstä.

Teoreettisissa tarkasteluissa tiheysfunktion oletetaan usein olevan mielivaltainen tai kuuluvan johonkin tunnettuun funktioperheeseen. Käytännön data-analyysissä puolestaan lähtökohtana on usein kokoelma havaintoja, joiden oletetaan olevan peräisin saman tiheysfunktion määräämästä todennäköisyysjakaumasta. Tällöin pyrkimyksenä on tehdä havaintojen perusteella arvio todellisesta, tuntemattomasta tiheysfunktiosta. Tällaista toimintaa kutsutaan tiheysfunktion estimoinniksi ja saatuja arvioita estimaateiksi.

Tilanteissa, joissa todellinen tiheysfunktio ei ole kovin siististi käyttäytyvä eli sileä funktio, ei estimoinnin tavoitteena välttämättä ole todellisen tiheysfunktion löytäminen. Varsinkin eksploratiivisten tarkastelujen yhteydessä on tällöin tarkoituksenmukaisempaa löytää todellisen tiheysfunktion pääpiirteet riittävän hyvin säilyttävä sileä approksimaatio eli silote. Tällaista toimintaa kutsutaan silottamiseksi (engl. *smoothing*). Vaikka estimoinnin ja silottamisen tulkitaankin usein tarkoittavan samaa asiaa, halutaan tässä työssä tehdä ero toimintojen kesken juuri niiden erilaisten päämäärien johdosta.

Estimoinnin lähestymistavasta riippuen estimointimenetelmät voidaan jakaa karkeasti parametrisiin ja parametrittomiin menetelmiin. Parametrisissa estimointimenetelmissä on tapana ensin kiinnittää aineiston sovitteeksi eli käytettäväksi silotteeksi jokin sileä funktio, jonka analyyttinen lauseke tunnetaan. Tämän jälkeen estimoidaan lausekkeen parametrit, joiden avulla hienosäädetään silotteen sijaintia ja muotoa. Jos puolestaan ensin kiinnitetään tai optimoidaan menetelmän parametrit ja annetaan silotteen määräytyä kokonaisuudessaan näiden menetelmäparametrien ja aineiston avulla, niin puhutaan, vaikkakin hieman ristiriitaisesti, parametrittomasta estimoinnista. Näin meneteltynä silotteen sileä tai rosoinen muoto syntyy ai-

neistolähtöisesti ja siihen voidaan vaikuttaa menetelmäparametrien avulla. Näin ollen parametrittomassa estimoinnissa ennakkotietämykselle tai oletuksille todellisen funktion muodosta ei ole tarvetta.

Vaikka parametrittomien estimointimenetelmien avulla voidaan tuottaa hyvin monenlaisia tuloksia, suositaan menetelmissä usein menetelmäparametrien automaattista optimointia ja vain yhden silotteen tuottamista menetelmän tulokseksi. P. Chaudhurin ja J. S. Marronin [1] kehittämä SiZer-menetelmä perustuu ajatukseen, ettei silotteen rosoisuudelle tai sileydelle ole olemassa yhtä ainoaa oikeaa vastausta. Heidän mukaansa jokainen silote edustaa sitä tietämystä, mikä on aineistosta saatavissa tarkasteltaessa tilannetta kyseisellä tarkkuudella eli resoluutiolla. SiZer-menetelmässä tarkastellaankin samanaikaisesti useita erilaisia parametrittoman estimoinnin tuottamia silotteita ja hyödynnetään tilastollista merkitsevyystestausta tehtäessä päätelmiä siitä, mitkä aineiston ennusteisiin synnyttämät piirteet kuuluvat todelliseen funktioon ja mitkä puolestaan aiheutuvat satunnaisvirheestä. Näin ollen SiZer-menetelmä yhdistää frekventistisen tilastollisen päättelyn tietokonenäön tutkimusalalla esiintyvään ideaan mikro- ja makroskooppisesta silotteen tarkastelutavasta ja tuo uuden mielenkiintoisen lähestymistavan tiheysfunktion estimointiin.

Tässä työssä käsitellään tiheysfunktion estimointia ja saatujen estimointitulosten analysointia SiZer-menetelmän avulla. Lisäksi SiZer-menetelmästä kehitetään tiheysfunktion estimointiin soveltuva versio, jossa tilastollinen päättely toteutetaan frekventistisen päättelyn sijaan Bayes-päättelyn avulla. Tämän toivotaan tehostavan perinteisen SiZer-menetelmän toimintaa esimerkiksi tilanteissa, joissa havaintoja on käytettävissä vähän joko kokonaisuudessaan tai vain aineiston tietyistä osista. Lisäksi työn toivotaan herättävän kiinnostusta SiZer-menetelmään myös Bayes-tilastotiedettä harjoittavien keskuudessa.

Työn loppuosa rakentuu siten, että luvussa 2 käsitellään tiheysfunktion estimointia ja esitellään muutamia keskeisiä estimointimenetelmiä. Luvussa 3 esitellään SiZer-menetelmän lähtökohdat, toimintaperiaatteet ja menetelmän tulokset eli SiZer-värikartat. Luvussa 4 esitellään Bayes-Logspline -menetelmä esimerkkinä Bayes-päättelyä hyödyntävästä tiheysfunktion estimointimenetelmästä ja kehitetään Bayes-Logspline -menetelmän avulla SiZer-menetelmästä Bayes-SiZer -menetelmä. Luvussa 5 testataan Bayes-SiZer -menetelmää synteettisten ja empiiristen aineistojen avulla sekä analysoidaan saatuja tuloksia. Tulosten perusteella tehdyt johtopäätökset esitellään luvussa 6.

2 Tiheysfunktion estimointi

Todennäköisyysavaruudessa $(\Omega, \mathcal{A}, \mathbb{P})$ määritellyllä satunnaismuuttujalla $X : \Omega \to \mathbb{R}$ sanotaan olevan jatkuva jakauma tiheysfunktiolla f, jos funktio $f : \mathbb{R} \to [0, \infty)$ on integroituva ja Borel-mitallinen sekä tapahtumien $\{X \in B\} = X^{-1}(B) \in \mathcal{A}$ todennäköisyydet voidaan ilmaista lausekkeella

$$\mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x$$

kaikilla Borel-joukoilla B. Tällöin todennäköisyysmitan $\mathbb{P}$ ominaisuudesta $\mathbb{P}(\Omega) = \mathbb{P}(X \in \mathbb{R}) = 1$ seuraa välittömästi, että jokainen tiheysfunktio f toteuttaa ehdon

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1\,. \tag{1}$$

Vastaavasti jokainen Borel-mitallinen, melkein kaikkialla ei-negatiivinen ja ehdon (1) täyttävä integroituva funktio $f : \mathbb{R} \to [0, \infty)$ on jonkin satunnaismuuttujan tiheysfunktio [6].

Oletetaan, että satunnaismuuttujalla X on jatkuva jakauma tiheysfunktiolla f, ja merkitään tätä lyhyesti $X \sim f$. Lisäksi oletetaan, että satunnaismuuttujat $X_1, \ldots, X_n$ muodostavat satunnaisotoksen satunnaismuuttujan X jakaumasta. Satunnaisotoksella tarkoitetaan tässä työssä, että $X_1, \ldots, X_n \sim f$ ja kaikki satunnaismuuttujat X_i, $i = 1, \ldots, n$ ovat keskenään riippumattomia. Tässä työssä myös aineisto ja otos ovat ekvivalentteja ilmaisuja satunnaisotokselle ja satunnaisotoksen komponentteja X_i, $i = 1, \ldots, n$ kutsutaan havainnoiksi. Satunnaisotoksen avulla on nyt mahdollista konstruoida tuntemattomasta tiheysfunktiosta f arvioita, joita tässä työssä merkitään yleisesti $\hat{f}_n$. Tällaista toimintaa kutsutaan tiheysfunktion estimoinniksi ja saatuja arvioita estimaateiksi. Estimaattoriksi puolestaan kutsutaan realisoituneen estimaatin $\hat{f}_n$ satunnaismuuttujavastinetta.

Johtuen tiheysfunktion keskeisestä asemasta todennäköisyysteoriassa myös tiheysfunktion estimointi liittyy moneen sovellusalaan ja erityisesti sovelluksiin, joissa käsitellään empiirisiä aineistoja. Lukuisten todennäköisyyslaskennallisten käyttökohteiden lisäksi tiheysfunktion estimointia käytetään esimerkiksi tilastollisessa hahmontunnistuksessa luokittimien konstruointiin [3], ryhmittelyanalyysissä ryhmien muodostamiseen [24] ja tilastollisessa testauksessa uskottavuusosamäärien laskemiseen [21]. Lisäksi aineistojen eksploratiivisissa tarkasteluissa tiheysfunktion estimaattien avulla pyritään usein selvittämään aineistoille ominaisia piirteitä, kuten jakaumien vinoutta sekä lokaalien maksimien eli moodien lukumääriä ja sijainteja.

Aineiston eksploratiivisissa tarkasteluissa tiheysfunktion estimoinnin tavoitteena on usein löytää sileä funktio, jonka voidaan ainakin uskoa noudattavan todellisen tiheysfunktion pääpiirteitä. Koska todellinen tiheysfunktio saattaa kuitenkin olla varsin rosoinen, voidaan saadut estimaatit tulkita todellisen tiheysfunktion silotteiksi, jotka on saatu silottamalla satunnaisvirhettä sisältävää aineistoa. Silottaminen ja funktion yleisen muodon etsintä voikin olla monissa tapauksissa tarkoituksenmukaisempaa kuin todellisen funktion tarkka estimointi. Regressio-ongelman tapauksessa Holmström ja Erästö antavat tästä hyvän esimerkin artikkelissaan [12], jossa he tutkivat ilmaston lämpötilan yleistä trendiä tuhansien vuosien ajalta.

Estimointi- ja silottamismenetelmiä on useita ja ne voidaan lähtöoletustensa perusteella jakaa karkeasti joko parametrisiin tai parametrittomiin menetelmiin. Seuraavissa aliluvuissa esitellään lyhyesti sekä parametrisen että parametrittoman tiheysfunktion estimoinnin pääpiirteitä ja keskeisiä menetelmiä.

2.1 Parametrinen tiheysfunktion estimointi

Parametrisissa menetelmissä lähtökohtana on parametriavaruus $\boldsymbol{\Theta} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, jonka jokaista alkiota $\boldsymbol{\theta}$ vastaa tiheysfunktio $f(\cdot\,;\boldsymbol{\theta}) : \mathbb{R} \to [0,\infty)$. Oletuksena on, että tuntematon tiheysfunktio f kuuluu avaruuden $\boldsymbol{\Theta}$ määräämään tiheysfunktioperheeseen $\mathcal{F} = \{f(\cdot\,;\boldsymbol{\theta}) \,|\, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ eli $f = f(\cdot\,;\boldsymbol{\theta_0})$ jollakin $\boldsymbol{\theta_0} \in \boldsymbol{\Theta}$. Tiheysfunktion f estimoinnissa yritetään nyt löytää parametrille $\boldsymbol{\theta_0}$ estimaatti $\hat{\boldsymbol{\theta}}_n$, jolloin lopulliseksi estimaatiksi saadaan $\hat{f}_n = f(\cdot\,;\hat{\boldsymbol{\theta}}_n)$.

Esimerkki parametrisesta tiheysfunktioperheestä on normaalijakauman tiheysfunktioiden perhe, jonka jäsenet ovat muotoa

$$f(x\,;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}\,.$$

Tällöin parametrivektori on $\boldsymbol{\theta} = (\mu,\sigma^2)$ ja parametriavaruus $\boldsymbol{\Theta} = \mathbb{R} \times (0,\infty) \subset \mathbb{R}^2$. Optimaalisen parametrivektorin $\hat{\boldsymbol{\theta}}_n$ löytämiseksi voidaan soveltaa esimerkiksi suurimman uskottavuuden estimointia (ks. [25]).

Yleisesti parametristen estimointimenetelmien ongelmana voidaan pitää oletusta $f \in \mathcal{F}$. Menetelmien käyttö ei kuitenkaan vaadi tarkkaa tietämystä funktiosta f, kunhan saatu estimaatti $\hat{f}_n$ ymmärretään todellisen funktion silotteeksi, jonka muoto on ennalta kiinnitetty. Esimerkiksi jokaisen yksiulotteisen aineiston tiheysfunktioksi voidaan sovittaa normaalijakauma, estimoida aineistosta parametrit μ sekä σ^2

ja tulkita syntyvä estimaatti aineiston yksihuippuiseksi silotteeksi. Tällaisten estimaattien käyttökelpoisuus riippuu kuitenkin aina asiayhteydestä.

2.2 Parametriton tiheysfunktion estimointi

Parametrittomassa estimoinnissa tuntemattomasta tiheysfunktiosta f tehtävät oletukset ovat selvästi lievempiä ja estimointi tapahtuu aineistolähtöisemmin kuin parametrisessa estimoinnissa. Terminologiasta huolimatta parametriton estimointimenetelmä sisältää usein lukuisia menetelmäparametreja. Erona parametriseen estimointiin on, ettei tuntemattoman funktion oleteta riippuvan äärellisestä parametriavaruudesta $\mathbf{\Theta} \subset \mathbb{R}^d$.

Parametrittomassa estimoinnissa kiinnitetään ensin menetelmäparametreille arvot. Riippuen estimoinnin tavoitteista menetelmäparametrien arvot määrätään joko subjektiivisen valinnan kautta tai käyttämällä automaattisia valintasääntöjä, kuten ristiinvalidointia (ks. esim. [8]). Automaattisten menetelmien käyttö on suosittua muun muassa tulosten vertailtavuuden kannalta sekä automatisoinnin tuoman kätevyyden kannalta tehtäessä lukuisia estimointeja, mutta niiden käyttö voi myös johtaa aineiston suppeaan analysointiin. Menetelmäparametrien kiinnityksen jälkeen käytettävä menetelmä tuottaa aineiston ja menetelmäparametrien avulla deterministisen estimaatin. Ratkaisevaa on kuitenkin se, että syntyvän silotteen sileyttä voidaan säädellä menetelmäparametrien avulla, jolloin silotteen muotoa ei tarvitse olettaa tunnetuksi tai kiinnittää ennalta.

Parametrittoman estimoinnin ongelmana on, että jatkuvalle funktiolle f tuotetut parametrittomat estimaattorit $\hat{f}_n$ ovat aina harhaisia. Matemaattisemmin ilmaistuna, jos $x \in \mathbb{R}$ ja $n \in \mathbb{N}$, niin kaikilla estimaattoreilla $\hat{f}_n$ on olemassa sellainen jatkuva tiheysfunktio $f : \mathbb{R} \to [0, \infty)$ että, jos $X_1, \ldots, X_n \sim f$ on satunnaisotos, niin $\mathbb{E}_f \hat{f}_n(x) \neq f(x)$ (Rosenblatt, 1956 [21] mukaan). Edellä merkinnän $\mathbb{E}_f$ mukainen odotusarvo lasketaan satunnaismuuttujien $X_1, \ldots, X_n$ yhteisjakauman suhteen. Harhaa eli systemaattista virhettä voidaan kuitenkin yrittää minimoida, mutta tällöin varianssi eli satunnaishajonta kasvaa. Varianssin minimointi puolestaan lisää harhaisuutta. Ongelmaa kutsutaankin nimellä ”bias-variance trade-off”. Parametristen menetelmien yhteydessä tehtävässä parametrien estimoinnissa ei tätä ongelmaa ole, ja usein parametrien hyviltä estimaattoreilta vaaditaankin harhattomuutta. Kun tarkasteltava funktioperhe $\mathcal{F}$ on riittävän ”suppea”, voidaan parametrisessa estimoinnissa muodostaa harhattomia estimaattoreita myös funktioille. Tämä on mah-

dollista esimerkiksi luvun 2.1 normaalijakauman tiheysfunktioiden perheen kohdalla (ks. [5]).

Seuraavaksi esitellään muutamia perinteisiä ja keskeisimpiä parametrittomia tiheysfunktion estimointimenetelmiä. Pääpaino menetelmien käsittelyssä on niiden eksploratiivisiin tarkasteluihin liittyvissä ominaisuuksissa, kuten menetelmäparametrien vaikutuksissa syntyviin silotteisiin. Esimerkkiaineistona käytetään 485 havainnosta koostuvaa Hidalgo-postimerkkiaineistoa. Yksiulotteiset havainnot ovat mittauksia vuosina 1872–1874 Meksikossa julkaistujen Miguel Hidalgo y Costillan kuvalla koristeltujen postimerkkien paksuuksista (mittayksikkönä millimetri). Aineiston tekee mielenkiintoiseksi se, että siinä ilmenee paljon hajontaa ja ryvästeisyyttä, mikä viittaa postimerkkipaperin useaan eri lähteeseen. Eksploratiivisten tarkastelujen tavoitteena onkin selvittää mahdollisten paperilähteiden lukumäärä. Aineisto tuli artikkelin [13] kautta tunnetuksi tilastotieteellisessä kirjallisuudessa ja se on ladattavissa Internetistä StatLib-arkistosta osoitteesta `http://lib.stat.cmu.edu/jcgs/`.

2.2.1 Histogrammi

Histogrammi on ehkäpä tunnetuin parametriton tiheysfunktioestimaattori. Sen toteutuksessa reaaliakseli jaetaan ensin tyypillisesti yhtä pitkiin luokkaväleihin (engl. *bin*) A_k esimerkiksi kaavalla

$$A_k = [\alpha_0 + kh, \alpha_0 + (k+1)h), \qquad k \in \mathbb{Z}\,, \tag{2}$$

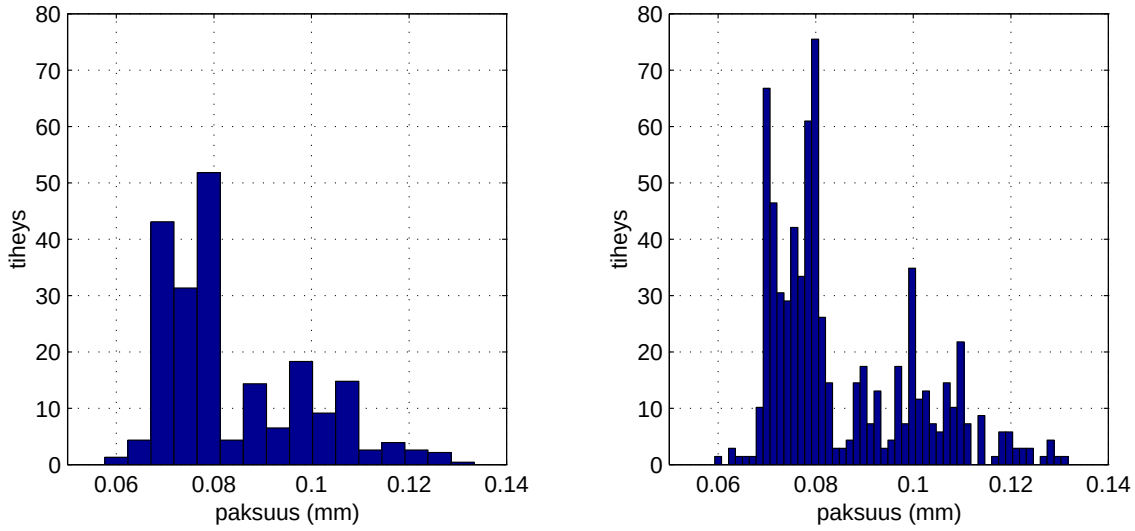
missä $h > 0$ on kunkin luokkavälin pituus (engl. *bin width*). Tämän jälkeen satunnaisotoksen $X_1, \ldots, X_n$ avulla voidaan muodostaa tiheysfunktiolle f estimaattori $\hat{f}_n(\cdot\,;\,h)$ lausekkeesta

$$\hat{f}_n(x\,;\,h) = \frac{1}{nh}\#\{i \mid X_i \in A_k, i = 1, \ldots, n\}\,, \qquad x \in A_k\,.$$

Realisoitunut estimaatti pisteessä x syntyy siis laskemalla kuinka moni havainnoista osuu samalle luokkavälille kuin evaluointipiste x itse.

Kuvassa 1 on kaksi Hidalgo-postimerkkiaineistosta muodostettua histogrammia, joissa on käytetty keskenään erisuuria luokkavälien pituuksia. Histogrammeista nähdään selvästi, miten luokkavälien pituus vaikuttaa estimaatin sileyteen; mitä suurempi luokkavälin pituus sitä sileämpi estimaatti. Menetelmäparametria h sanotaankin usein silotusparametriksi.

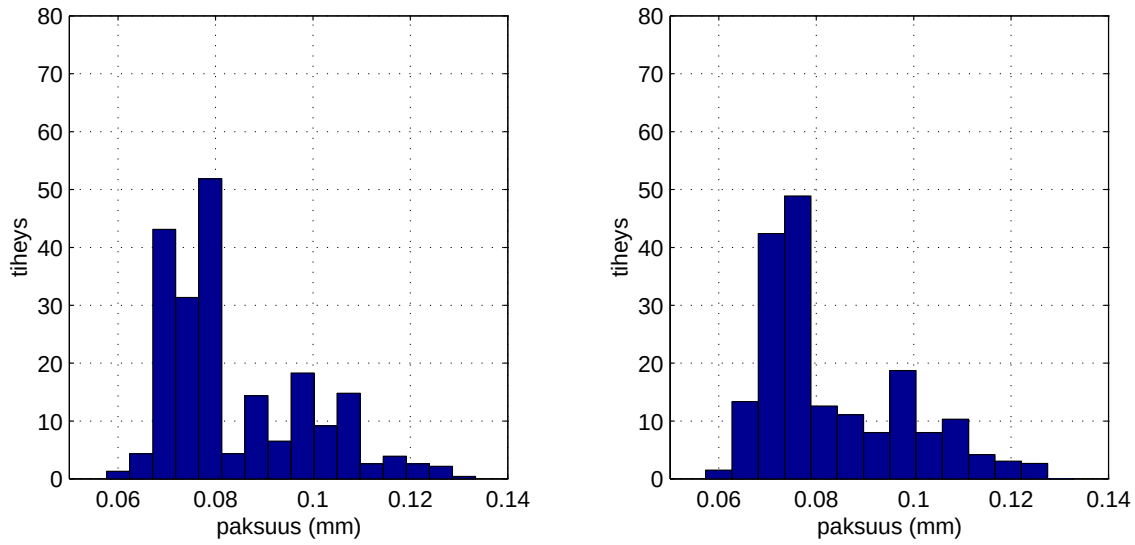
Luokkavälien pituuden lisäksi myös luokkavälien sijainti vaikuttaa muodostettavaan histogrammiin. Kaavassa (2) luokkavälin A_k sijaintia voidaan säädellä parametrin α_0 avulla. Kuvassa 2 on kaksi Hidalgo-postimerkkiaineistosta muodostettua histogrammia, joissa luokkavälien pituudet ovat samat, mutta luokkavälien sijainnit poikkeavat toisistaan. Vaikka sijaintien ero on vähäinen, antavat histogrammit varsin erilaiset estimaatit tuntemattomalle tiheysfunktolle ja paperilähteiden lukumäärälle. Vasemmanpuoleisen histogrammin moodeista voitaisiin päätellä, että mahdollisia paperilähteitä on peräti kuusi, kun taas oikeanpuoleisen histogrammin perusteella niitä näyttäisi olevan vain kolme.



Kuva 1. Kaksi Hidalgo-postimerkkiaineistosta muodostettua histogrammia eri luokkavälien pituuksilla.

Mikä kuvien 1 ja 2 histogrammeista sitten kertoo tuntemattomien paperilähteiden lukumäärän? Koska menetelmäparametrit yhdessä aineiston kanssa määrittelevät histogrammin, yhtä hyvin voidaan kysyä, mitkä ovat menetelmäparametrien optimaaliset arvot. Kuvien histogrammeja muodostettaessa sijainti- ja silotusparametrit on valittu subjektiivisesti, mutta etenkin silotusparametrin h valitsemiseksi on olemassa monia teoreettisiin tuloksiin nojaavia sääntöjä sekä aineistoa hyödyntäviä menetelmiä. Scott esittelee teoksessaan [21] muun muassa Sturgesin säännön ei-tyhjien luokkavälien lukumäärälle

$$K = 1 + \log_2 n \,,$$



Kuva 2. Kaksi Hidalgo-postimerkkiaineistosta muodostettua histogrammia samoilla luokkavälien pituuksilla mutta eri luokkavälien sijainneilla.

normaalijakaumaan perustuvan säännön luokkavälin pituudelle

$$h^* = 3.5\hat{\sigma} n^{-1/3}\,,$$

missä $\hat{\sigma}$ on otoskeskihajonta, sekä asymptoottisen keskimääräisen integroidun neliöllisen virheen (AMISE) minimoivan ratkaisun

$$h^* = \left(\frac{6}{nR(f')}\right)^{1/3},$$

missä funktion $g \in L^2(\mathbb{R})$ rosoisuutta edustaa termi

$$R(g) = \int_{\mathbb{R}} g(x)^2 \,\mathrm{d}x\,. \tag{3}$$

Aineistolähtöisistä menetelmistä mainittakoon harhaton ristiinvalidointi (engl. *Unbiased Cross-Validation*), jossa minimoidaan silotusparametrin h suhteen lauseketta

$$UCV(h) = R(\hat{f}_n(\cdot\,;\,h)\,) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i,n}(X_i\,;\,h)\,, \tag{4}$$

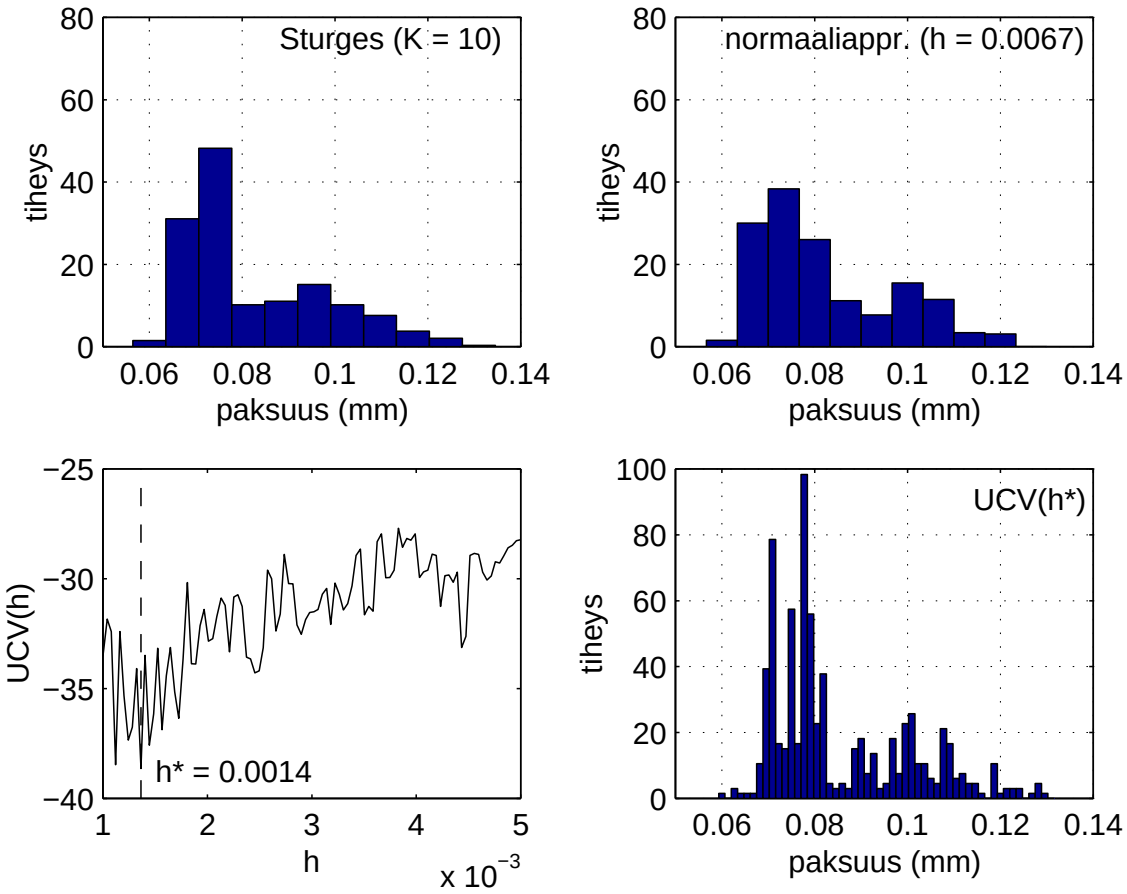
missä termi $\hat{f}_{-i,n}(X_i\,;\,h)$ vastaa ilman havaintoa X_i muodostettua histogrammia. Koska $UCV(h)$ on termin

$$\mathbb{E}\int_{\mathbb{R}} \Big[\hat{f}_n(x\,;\,h) - f(x)\Big]^2 \,\mathrm{d}x - R(f) = \mathrm{MISE}(h) - R(f)$$

harhaton estimaattori, lausekkeen (4) minimointi on ekvivalenttia keskimääräisen integroidun neliöllisen virheen MISE(h) minimoinnin kanssa (ks. [21]).

Kuvaan 3 on koottu Sturgesin säännön, normaaliapproksimaation ja harhattoman ristiinvalidoinnin tuottamat histogrammit. Lisäksi kuvan 3 vasemmassa alanurkassa on esitetty estimaattori (4) silotusparametrin h funktiona ja tämän minimoiva ratkaisu h^*. Selvästi nähdään, että Sturgesin sääntöön ja normaaliapproksimaatioon perustuvat histogrammit tuottavat lähes samankaltaiset silotetut kaksihuippuiset histogrammit. Ristiinvalidointi puolestaan tuottaa hyvin rosoisen histogrammin, jossa aineiston voidaan havaita keskittyvän arvojen 0.07 mm, 0.08 mm, 0.09 mm, 0.10 mm, 0.11 mm, 0.12mm ja 0.13 mm ympärille. Ristiinvalidoinnin tuottaman tuloksen puolesta puhuu Izenmanin ja Sommerin artikkelissa [13] saamat tulokset. He käyttivät ristiinvalidoinnin sijaan Silvermanin [24] kehittämää tilastollista testiä moodien lukumäärän selvittämiseksi ja saivat viitteitä seitsemän moodin olemassaolosta. Sturgesin säännön ja normaaliapproksimaation soveltamista Hidalgo-postimerkkiaineistoon voidaan kritisoida siitä, että nämä eivät pysty ottamaan huomioon aineiston tiheysfunktiossa selvästi ilmenevää vinoutta eli pitkää oikeaa häntää.

Edellisten tarkastelujen valossa optimaalisen histogrammin löytäminen ei selvästikään ole helppo tehtävä. Tarkasteltujen menetelmien lisäksi histogrammista on olemassa myös adaptiivisia versioita, joissa luokkavälien pituuksien sallitaan vaihdella (ks. [21]). Adaptiiviset menetelmät muistuttavat läheisesti χ^2-yhteensopivuustestiä, jossa arvoalueita eli soluja usein yhdistellään, jotta solut sisältäisivät riittävän määrän havaintoja. Lisäksi yksiulotteinen histogrammi voidaan yleistää useampaan ulottuvuuteen, missä luokkavälien sijaan tarkastellaan luokkalaatikoita ja luokkavälien pituuksien sijaan luokkalaatikoiden tilavuuksia. Useammassa ulottuvuudessa histogrammien käyttö asettaa kuitenkin suuria vaatimuksia laskentakapasiteetille. Vaikka histogrammi on yksinkertainen ja tehokas estimaattori, sen ongelmana on, että se on funktiona epäjatkuva luokkavälien reunoilla. Tämä vaikeuttaa histogrammin käyttöä esimerkiksi tässä työssä esiteltävien SiZer-menetelmien tapauksissa, joissa ollaan kiinnostuneita ennustettavan tiheysfunktion derivaatoista.



Kuva 3. Ylärivissä Sturgesin sääntöön ja normaaliapproksimaatioon perustuvat histogrammit ja alarivissä harhattoman ristiinvalidoinnin minimiratkaisu (merkitty kuvaan pisteviivalla) ja siihen perustuva histogrammi.

2.2.2 Ydinestimointi

Histogrammin idea voidaan yleistää niin kutsutuksi naiiviksi estimaattoriksi

$$\hat{f}_n(x\,;\,h) = \frac{1}{2hn}\#\{i \,|\, X_i \in (x-h,\, x+h], i = 1, \ldots, n\}\,, \tag{5}$$

joka voidaan kirjoittaa myös muotoon

$$\hat{f}_n(x\,;\,h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left(\frac{x - X_i}{h}\right)\,, \tag{6}$$

missä painofunktio $w : \mathbb{R} \to \mathbb{R}$ on

$$w(x) = \begin{cases} 1/2, & \text{kun } -1 \le x < 1 \\ 0, & \text{muuten} \end{cases}\,.$$

Naiivi estimaattori vastaa siis histogrammia, mikäli evaluointipiste x sijaitsee keskellä histogrammin luokkaväliä, jonka pituus on $2h$. Kuten histogrammi ei naiivi estimaattorikaan ole jatkuva funktio, vaan funktio hyppää jokaisessa pisteessä $X_i \pm h$ derivaatan ollessa nolla kaikissa muissa pisteissä. Tästä johtuen myös naiivin estimaattorin tuottamat estimaatit ovat laatikkomaisia porrasfunktioita, joiden kuvallinen esitys ei ole täysin tyydyttävä. Erona histogrammiin on kuitenkin histogrammin sijaintiparametrin α_0 puuttuminen. Näin ollen ainoastaan silotusparametri h vaikuttaa naiivin estimaattorin tuottamien estimaattien sileyteen.

Myös naiivi estimaattori on mahdollista yleistää joustavammaksi menetelmäksi. Korvaamalla naiivin estimaattorin (6) painofunktio w yleisemmällä funktiolla $K : \mathbb{R} \to \mathbb{R}$ eli ytimellä, joka niin ikään toteuttaa ehdon

$$\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1\,, \tag{7}$$

voidaan muodostaa funktion f ydinestimaattori

$$\hat{f}_n(x\,;\,h) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)\,.$$

Merkitsemällä silotusparametrilla skaalattua ydintä

$$K_h(t) = \frac{1}{h} K\left(\frac{t}{h}\right)\,, \qquad t \in \mathbb{R}$$

voidaan ydinestimaattori ilmaista skaalattujen ydinten aritmeettisena keskiarvona

$$\hat{f}_n(x\,;\,h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\,. \tag{8}$$

Estimoitaessa tiheysfunktiota kannattaa ytimeksi valita jokin tiheysfunktio. Tällöin ydinestimaattori $\hat{f}_n(\cdot\,;\,h)$ on ei-negatiivinen, integroituva sekä toteuttaa ehdon

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{f}_n(x\,;\,h)\,\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K_h(x - X_i)\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K(y)\,\mathrm{d}y \qquad \text{muuttujan vaihto } y = (x - X_i)/h \\
&= 1\,.
\end{aligned}$$

Siis myös ydinestimaattori $\hat{f}_n(\cdot\,;\,h)$ on tiheysfunktio. Koska ydinestimaattori (8) perii myös ytimen jatkuvuus- ja derivoituvuusominaisuudet, on yhdeksi suosituimmista ytimistä noussut kaikkialla jatkuva ja mielivaltaisen useasti derivoituva standardinormaalijakauman $N(0,1)$ tiheysfunktio eli Gaussin ydin

$$K(x) = (\sqrt{2\pi})^{-1} \exp\{-x^2/2\}\,, \quad x \in \mathbb{R}\,. \tag{9}$$

Gaussin ydin soveltuu lisäksi erityisen hyvin moodien etsintään, sillä se toteuttaa seuraavan lauseen, jonka Silverman todisti artikkelissaan [23].

Lause 1 *Oletetaan, että $X_1, \ldots, X_n$ on mielivaltainen kokoelma havaintoja ja $\hat{f}_n(\cdot\,;\,h)$ kaavassa (8) määritelty ydinestimaattori Gaussin ytimellä K. Tällöin jokaisella kiinnitetyllä ei-negatiivisella kokonaisluvulla m funktion $\partial^m \hat{f}_n(x\,;\,h)/\partial x^m$ maksimien lukumäärä x:n vaihdellessa on oikealta jatkuva, parametrin h suhteen laskeva funktio.*

Lauseen 1 seurauksena siis Gaussin ydintä hyödyntävän ydinestimaattorin aineistosta löytämät moodit käyttäytyvät monotonisesti silotusparametrin h suhteen. Tulokseen palataan SiZer-menetelmän yhteydessä luvussa 3.

Ydinestimaattorin odotusarvoksi jokaisessa pisteessä x saadaan

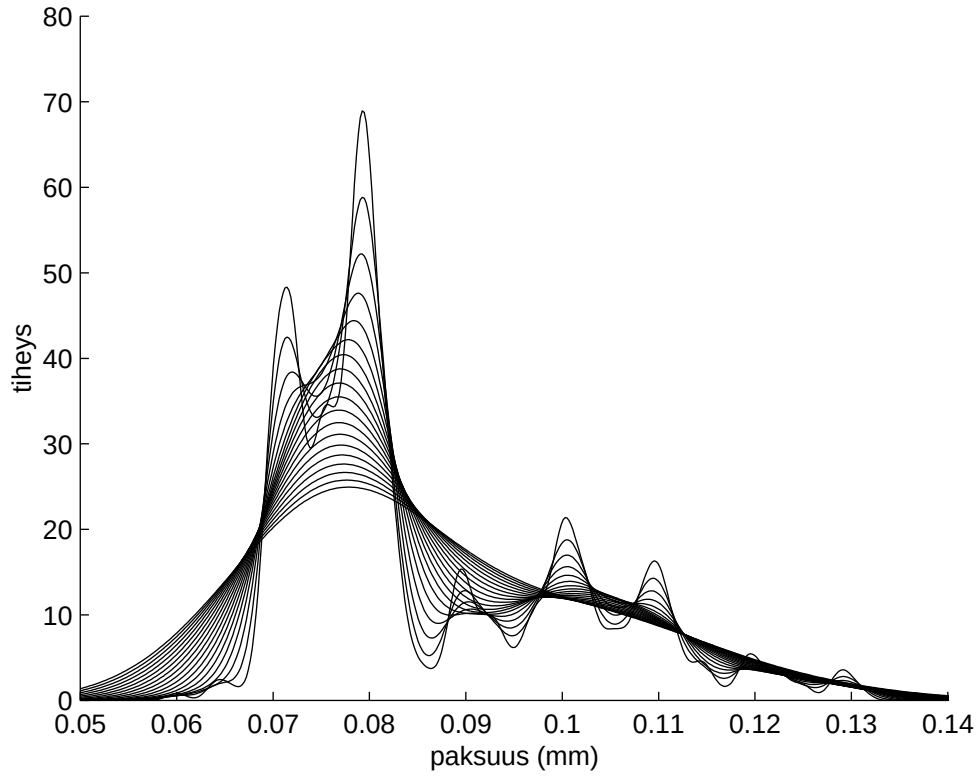
$$\mathbb{E}(\hat{f}_n(x\,;\,h)) = \mathbb{E}(K_h(x - X)) = \int_{-\infty}^{\infty} K_h(x-y) f(y)\,\mathrm{d}y =: (f * K_h)(x) \tag{10}$$

eli ydinestimaattorin odotusarvo on itse asiassa todellisen tiheysfunktion ja skaalatun ytimen konvoluution tuottama silote. Olettamalla, että ydin K toteuttaa sellaiset säännöllisyysehdot, että x:n suhteen derivoinnin ja y:n suhteen integroinnin järjestystä voidaan vaihtaa, pätee vastaavasti

$$\mathbb{E}(\hat{f}_n'(x\,;\,h)) = (f * K_h')(x) = (f * K_h)'(x)\,. \tag{11}$$

Näin ollen ydinestimoinnilla on mielenkiintoinen yhtymäkohta konvoluutiolla silottamiseen. Sovellusaloja, joissa hyödynnetään konvoluutiolla silottamista ovat esimerkiksi digitaalinen kuvankäsittely ja tietokonenäkö. Edellisessä konvoluutiota käytetään suodattamaan eli silottamaan kohinaa signaaleista (ks. [16]) ja jälkimmäisessä konvoluution avulla suodatetaan kuvien yksityiskohtia (ks. [22]). Ydinestimoinnin ja konvoluutiolla silottamisen väliseen yhteyteen palataan luvussa 3, missä tällä yhteydellä on keskeinen merkitys SiZer-menetelmän ajatusmaailmassa.

Kuvassa 4 on esitetty Hidalgo-postimerkkiaineistosta muodostettuja ydinestimaatteja eri silotusparametrin h arvoilla käyttäen Gaussin ydintä. Kuten histogrammin tapauksessa, myös nyt silotusparametri h vaikuttaa estimaatin muotoon siten, että mitä suurempi on parametrin h arvo, sitä sileämpi on estimaatti. Tarkastelemalla kuvan 4 ydinestimaattien perhettä havaitaan postimerkkiaineiston moodien lukumäärän olevan välillä 1–7 silotusparametrin suuruudesta riippuen.



Kuva 4. Ydinestimaatteja eri silotusparametrin h arvoilla. Ytimenä on käytetty Gaussin ydintä.

Ydinestimointia ja sen matemaattisia ominaisuuksia on tutkittu paljon. Kuten histogrammin silotusparametrille, on myös ydinestimaattorin silotusparametrille olemassa valintasääntöjä ja -menetelmiä. Samoin tavalliselle ydinestimoinnille on myös kehitetty adaptiivisia muunnelmia, joissa silotusparametrin sallitaan muuttuvan estimoitavan alueen mukana. Lisäksi ydinestimointia voidaan soveltaa myös moniulotteiseen tilanteeseen. [21, 24]

Muita yleisiä parametrittomia estimointimenetelmiä ovat lähinaapurimenetelmä, ortogonaalisarjaestimaattori sekä sakotetun uskottavuuden menetelmä. Näiden mene-

telmien tarkastelu jätetään kuitenkin tämän työn ulkopuolelle. Seuraavassa luvussa esitellään alkuperäinen frekventistiseen tilastolliseen päättelyyn ja ydinestimointiin perustuva SiZer-menetelmä.

3 SiZer-menetelmä

Kuten luvussa 2 havaittiin, silottaminen vaikuttaa suuresti estimaatista havaittaviin piirteisiin. Vähäinen silottaminen saattaa jättää estimaattiin piirteitä, jotka ovat syntyneet ainoastaan satunnaisvirheen johdosta, ja liika silottaminen voi estää havainnoimasta todellisia piirteitä. Etsittäessä satunnaisvirhettä sisältävän aineiston todellisia piirteitä onkin hyvä tarkastella useita eri silotteita ja tutkia piirteiden ja silotteiden riippuvuuksia. Tämänkaltainen silotteiden perhettä tutkiva lähestymistapa on lähtökohtana myös Chaudhurin ja Marronin [1] kehittämässä SiZer-menetelmässä.

SiZer-menetelmä on kehitetty eksploratiivisen data-analyysin avuksi ja se soveltuu tiheysfunktion estimoinnin lisäksi myös regressiofunktion estimointiin. Regressiofunktion estimoinnin tapauksessa SiZer-menetelmä käyttää estimointiin lokaalia lineaarista regressiota [8] ja tiheysfunktion estimointiin ydinestimointia Gaussin ytimellä. Tässä työssä keskitytään SiZer-menetelmän osalta ainoastaan tiheysfunktion estimointiin, mutta esitettävät periaatteet soveltuvat asianmukaisesti muutettuna myös regressiotapaukseen. SiZer-menetelmän MATLAB-toteutus on ladattavissa Marronin kotisivuilta `http://www.stat.unc.edu/faculty/marron.html`.

Menetelmässä tarkastellaan samanaikaisesti tuntemattoman funktion f useita eri silotteita $f * K_h$ ja kohdistetaan tilastollinen päättely tuntemattoman funktion itsensä sijaan sen silotteisiin. Ajatuksena on, että jokainen silote $f * K_h$ sisältää kaiken aineistosta saatavilla olevan informaation, kun aineistoa tarkastellaan kiinnitetyn silotusparametrin h määräämällä tarkkuudella eli resoluutiolla. Tämä on tavallinen lähestymistapa tietokonenäön tutkimusalalla, jonka mallien mukaisesti parametrin h suuret arvot mallintavat kuvien makroskooppista eli kaukaa katsottua tarkastelua ja pienet arvot mikroskooppista eli läheltä katsottua tarkastelua [16]. Menetelmä eroaa siis perinteisestä estimoinnista, jossa tilastollisen päättelyn kohteena on tuntematon funktio f itse.

3.1 Toimintaperiaate

Oletetaan, kuten luvussa 2, että $X \sim f$ ja satunnaismuuttujat $X_1, \ldots, X_n$ muodostavat satunnaisotoksen satunnaismuuttujan X jakaumasta. Olkoon $\hat{f}_n(\cdot\,;\,h)$ kaavan (8) mukainen tuntemattoman tiheysfunktion f ydinestimaattori skaalatulla Gaussin ytimellä K_h. Tällöin tuloksen (10) nojalla pätee $\mathbb{E}(\hat{f}_n(\cdot\,;\,h)) = f * K_h$, jolloin $\hat{f}_n(\cdot\,;\,h)$ on silotteen $f * K_h$ harhaton estimaattori. Vastaavasti tuloksen (11) nojalla tiheysfunktion derivaatan ydinestimaattori $\hat{f}'_n(\cdot\,;\,h)$ on silotteen derivaatan $(f * K_h)'$ harhaton estimaattori.

SiZer-menetelmässä etsitään todellisia ja tilastollisesti merkitseviä piirteitä, kuten funktion lokaaleja minimi- ja maksimiarvoja, tarkastelemalla funktion derivaatan nollakohtia. Nimi SiZer onkin lyhennys englanninkielisistä sanoista *SIgnificant ZERo crossings of the derivative*. Piirteen sanotaan olevan merkitsevä mikäli piirteen molemmin puolin derivaatta eroaa tilastollisesti merkitsevästi nollasta ja on eri puolilla vastakkaista merkkiä (+/-). Kun tilastollinen päättely kohdistetaan todellisen tiheysfunktion derivaatan f' sijaan silotteen derivaattaan $(f * K_h)'$, vältytään käyttämästä parametrittomalle estimoinnille ominaisia harhaisia estimaattoreita (vrt. luku 2.2).

Tilastolliset päättelyt derivaatoista tehdään luottamusvälien avulla. Koska estimaattori $\hat{f}'_n(\cdot\,;\,h)$ on harhaton, ovat muotoa

$$\left[\hat{f}'_n(x\,;\,h) - q \cdot s(\hat{f}'_n(x\,;\,h)), \hat{f}'_n(x\,;\,h) + q \cdot s(\hat{f}'_n(x\,;\,h))\right] \tag{12}$$

olevat symmetriset luottamusvälit keskittyneitä 'oikean' arvon ympärille. Luottamusvälin lausekkeessa (12) termi q edustaa fraktiilipistettä ja s keskihajontaa. Luottamusvälit voidaan muodostaa kiinteällä silotusparametrin h arvolla joko pisteittäisesti jokaiselle arvolle x tai samanaikaisesti yli kaikkien arvojen x. On myös mahdollista muodostaa luottamusvälit samanaikaisesti yli kaikkien arvojen h ja x. Luottamusvälien toteutuksessa fraktiilipisteen q valinta perustuu joko normaaliapproksimaatioon tai Bootstrap-menetelmään. Tarkastelu rajoitetaan tässä vain normaaliapproksimaation hyödyntämiseen pisteittäisissä ja pisteiden x suhteen samanaikaisissa luottamusväleissä.

Normaaliapproksimaation käyttö on keskeisen raja-arvolauseen nojalla perusteltua, sillä ydinestimaattorin derivaatta $\hat{f}'_n(\cdot\,;\,h)$ on samoin jakautuneiden satunnaismuuttujien $K'_h(x - X_i)$ aritmeettinen keskiarvo (vrt. kaava (8)). Olkoon $\Phi : \mathbb{R} \to [0, 1]$

standardinormaalijakauman kertymäfunktio ja $1-\alpha$ luottamustaso, kun $\alpha \in [0,1]$. Tällöin pisteittäisten luottamusvälien normaaliapproksimaatioksi saadaan

$$1-\alpha = \mathbb{P}\left(\left|\frac{\hat{f}_n'(x\,;\,h) - \mathbb{E}(\hat{f}_n'(x\,;\,h))}{s(\hat{f}_n'(x\,;\,h))}\right| \leq q_1\right) \approx 2\Phi(q_1) - 1\,,$$

jolloin fraktiilipisteeksi q_1 tulee

$$q_1 \approx \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\,.$$

Samanaikaisten luottamusvälien laskennasssa hyödynnetään etäällä toisistaan sijaitsevissa pisteissä x_1 ja x_2 evaluoitujen ydinestimaattorin arvojen $\hat{f}_n'(x_1\,;\,h)$ ja $\hat{f}_n'(x_2\,;\,h)$ approksimatiivista riippumattomuutta. Tällaisissa pisteissä ydinestimaattorin arvojen voidaan tulkita olevan lähes riippumattomia, sillä kaukana toisistaan sijaitsevien ydinten vaikutus syntyvän ydinestimaatin arvoon on pieni (vrt. kaava (8)). Samanaikainen luottamusväli voidaan näin ollen tuottaa approksimoimalla väliä äärellisellä määrällä riippumattomia luottamusvälejä, kun riippumattomien osavälien lukumäärä $m(h)$ estimoidaan aineistosta erikseen jokaiselle silotusparametrille h. Merkitään osaväliä $j = 1, \ldots, m(h)$ edustavaa ydinestimaattoria $\hat{f}_j'(\cdot\,;\,h)$. Tällöin riippumattomuuden nojalla samanaikaisten luottamusvälien normaaliapproksimaatioksi saadaan

$$\begin{aligned}
1-\alpha &= \prod_{j=1}^{m(h)} \mathbb{P}\left(\left|\frac{\hat{f}_j'(x\,;\,h) - \mathbb{E}(\hat{f}_j'(x\,;\,h))}{s(\hat{f}_j'(x\,;\,h))}\right| \leq q_2\right) \\
&= \left[\mathbb{P}\left(\left|\frac{\hat{f}_1'(x\,;\,h) - \mathbb{E}(\hat{f}_1'(x\,;\,h))}{s(\hat{f}_1'(x\,;\,h))}\right| \leq q_2\right)\right]^{m(h)} \\
&\approx (2\Phi(q_2) - 1)^{m(h)}\,,
\end{aligned}$$

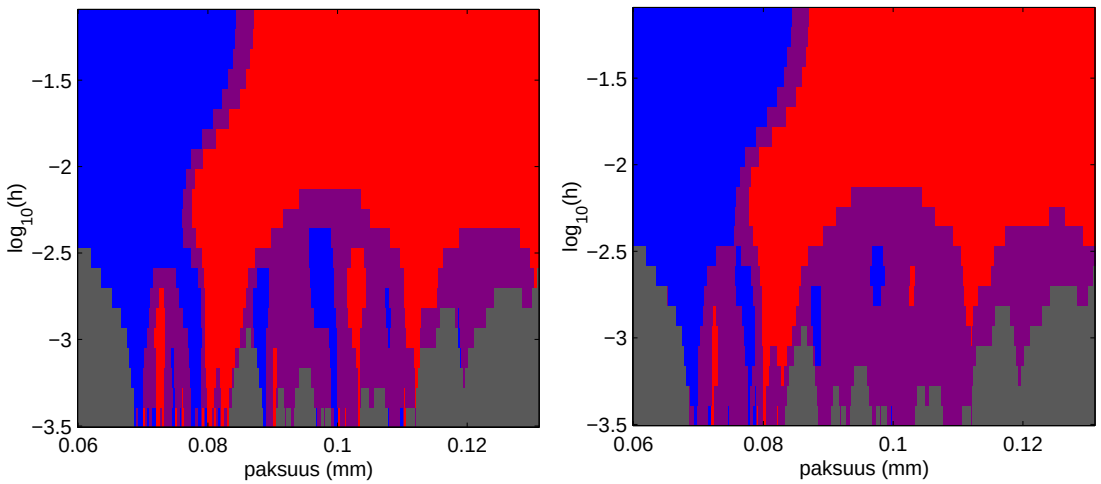
jolloin fraktiilipisteeksi q_2 tulee

$$q_2 \approx \Phi^{-1}\left(\frac{1 - (1-\alpha)^{1/m(h)}}{2}\right)\,.$$

Ydinestimaattorin lausekkeessa (8) olevan summan evaluointi tapahtuu SiZer-menetelmän toteutuksessa toistuvasti, mikä tekee SiZer-menetelmästä laskentaintensiivisen. Luottamusvälien (12) keskihajonnan $s(\hat{f}_n'(x\,;\,h))$ laskennan tehostamiseksi SiZer-menetelmässä käytetään aineiston lineaarista uudelleenryhmittelyä osaväleihin (engl. *linear binning*) [8]. Tämä nopeuttaa laskentaa mutta lisää menetelmään ylimääräistä approksimointia.

3.2 SiZer-värikartat

Tuloksenaan SiZer-menetelmä ei tuota varsinaista estimaattia tiheysfunktiosta, vaan kuvan 5 kaltaisia SiZer-värikarttoja, joista on nähtävissä tilastollisen päättelyn tuottamat tulokset. Sinisellä värillä on merkitty ne pisteet (x, h), joissa derivaatta $\hat{f}_n'(x\,;\,h)$ on merkitsevästi positiivinen, ja vastaavasti punaisella värillä ne pisteet (x, h), joissa $\hat{f}_n'(x\,;\,h)$ on merkitsevästi negatiivinen. Lilalla värillä puolestaan on merkitty pisteet, joissa $\hat{f}_n'(x\,;\,h)$ ei eroa merkitsevästi nollasta. Harmaa alue koostuu niistä pisteistä, joiden kohdalla menetelmä ei pysty tekemään luotettavia päätelmiä derivaatasta. Näin käy esimerkiksi silloin, kun fraktiilin laskennassa muodostetuvalle osavälille jää normaaliapproksimaation laskentaa varten liian vähän havaintoja.



Kuva 5. Hidalgo-postimerkkiaineistosta tuotetut SiZer-värikartat 95 %:n luottamustasolla. Vasemmanpuoleisessa kartassa luottamusvälit on laskettu pisteittäisesti ja oikeanpuoleisessa samanaikaisesti yli x-akselin pisteiden.

Mikäli pisteittäiset tai samanaikaiset luottamusvälit on laskettu kiinteällä silotusparametrin h arvolla, luetaan SiZer-värikarttaa tarkastelemalla yhtä pystyakselin eli logaritmoidun silotusparametrin $\log_{10}(h)$ arvoa kerrallaan. Tällöin värien peräkkäiset yhdistelmät kiinnitetyllä tasolla kertovat ne kohdat, joissa piirteet ovat merkitseviä. Esimerkiksi kuvan 5 Hidalgo-postimerkkiaineistosta tuotetuissa SiZer-värikartoissa merkitsevät moodit sijaitsevat kohdissa, joissa sininen väri vaihtuu punaiseksi eli derivaatan kasvu muuttuu laskuksi. Näin käy vasemmanpuoleisessa kartassa likimain x-akselin kohdissa 0.07 mm, 0.08 mm, 0.10 mm ja 0.11 mm, kun karttaa tarkastellaan arvolla $\log_{10}(h) = -3$.

Kun kuvan 5 vasemmanpuoleista SiZer-värikarttaa verrataan oikeanpuoleiseen karttaa, nähdään selvästi, miten tarkemmilla resoluutioilla lilan värin osuus kasvaa oikeanpuoleisessa kartassa. Tämä johtuu juuri samanaikaisten luottamusvälien konservatiivisuudesta suhteessa pisteittäisiin luottamusväleihin. Chaudhuri ja Marron eivät kuitenkaan suosittele käytettäväksi pisteittäisiä luottamusvälejä, sillä tällöin tulokseksi saatetaan saada liikaa merkitseviä piirteitä. Samanaikaisten luottamusvälien tuottamat kartat ovat lisäksi tulkinnallisesti helpompia, sillä käytettäessä pisteittäisiä luottamusvälejä voidaan päätelmät luottamustasolla $1 - \alpha$ tehdä ainoastaan lokaalisti eikä koko x-akselin suhteen.

Kuvan 5 SiZer-värikartoista nähdään myös Gaussin ytimen ja lauseen 1 vaikutus silotteiden piirteisiin. Merkitsevien moodien lukumäärän huomataan laskevan silotusparametrin kasvaessa. Näin tapahtuu, sillä lauseen 1 tulos pätee myös odotusarvon derivaatoille $\partial^s \mathbb{E}\hat{f}_n(\cdot\,;\, h)/\partial x^s$, $s = 0, 1, 2, \ldots$, minkä Marron ja Chaudhuri todistavat regressioestimaattorien tapauksessa artikkelissa [2]. Nämä tulokset ovat tärkeitä, sillä tarkasteltaessa erilaisten silotteiden joukkoa edeten runsaasta silottamisesta vähäiseen silottamiseen on menetelmien kannalta suotavaa, että piirteet häviävät monotonisesti. Jollei näin olisi, vaikeuttaisi se sopivien silotteiden löytämistä sekä silotteiden ja piirteiden riippuvuuksien tarkastelua. Piirteiden monotonisuuden johdosta myös SiZer-värikarttojen tulkitseminen ja johtopäätösten teko on helpompaa.

Kun kuvan 5 oikeanpuoleista karttaa tarkastellaan tasolla $\log_{10}(h) = -3$ ja verrataan luvussa 2.2 saatuihin tuloksiin, voidaan nähdä, että SiZer on kuvan 4 ydinestimaattien kanssa samaa mieltä moodeista kohdissa 0.07 mm, 0.08 mm ja 0.10 mm. Moodille kohdassa 0.09 mm SiZer tunnistaa merkittävän nousun, mutta ei tunnista merkittävää laskua. Vastaavasti ydinestimaattien tunnistamalle moodille kohdassa 0.11 mm SiZer tunnistaa ainoastaan merkittävän laskun. Kohdissa 0.12 mm ja 0.13 mm oleville moodeille SiZer ei pysty tuottamaan luotettavia päätelmiä, jonka johdosta alueet värjätään harmaalla.

3.3 Menetelmän vahvuudet ja heikkoudet

SiZer-värikartoista ei siis ilmene vain merkittävien piirteiden lukumäärät vaan myös niiden sijainnit. Juuri moodien etsinnässä (engl. *bump hunting*) tämä on parannus muihin menetelmiin, jotka perustuvat ainoastaan moodien lukumääriä koskevien hypoteesien tilastolliseen testaukseen (ks. [24]). Lisäksi SiZer-värikartat mahdollistavat päätelmien teon useasta silotteesta yhden kuvan avulla. Tällöin tutkittavasta

ilmiöstä saadaan laajempi kuva kuin tarkasteltaessa vain yhtä jossakin mielessä optimaalista silotetta.

Mikäli havaintopisteitä on aineistossa vähän tai harvasti, ei luottamusvälejä ole mielekästä konstruoida normaaliapproksimaation avulla. Tällöin SiZer ei pysty tekemään luotettavia päätelmiä ja kyseiset kohdat värjätään SiZer-värikartoissa harmaiksi. Kuvan 5 SiZer-värikartat ovat hyviä esimerkkejä siitä, miten harmaata väriä esiintyy usein paljon aineiston reunoilla äärimmäisten havaintojen vähäisyyden johdosta. Tiheysfunktion estimoinnin kannalta tämä on harmillista, sillä esimerkiksi riskiteoreettisissa sovelluksissa kiinnostuksen kohteena on usein juuri jakaumien häntien käyttäytyminen.

3.4 Menetelmän kehittäminen

Tässä työssä pyritään jatkokehittämään SiZer-menetelmää tiheysfunktion estimoinnin osalta. Edellä kuvattujen SiZer-menetelmän heikkouksien motivoimana tässä työssä on kehitetty tiheysfunktion estimointiin soveltuva Bayes-päättelyyn perustuva versio SiZer-menetelmästä. Bayes-päättelyn avulla on mahdollista välttää normaaliapproksimaatioon liittyvät ongelmat ja tehdä päätelmiä myös pienillä tai harvoilla aineistoilla. Idea ei ole uusi, sillä regressiofunktion estimointiin soveltuvan Bayes-SiZer -menetelmän eli BSiZerin ovat kehittäneet Erästö ja Holmström artikkelissaan [7]. Tässä työssä esiteltävä Bayes-SiZer -menetelmä perustuukin paljolti heidän työhönsä ja tuloksiinsa. Regressio- ja tiheysfunktion estimointiongelmien erilaisuuden johdosta tässä työssä esiteltävä Bayes-SiZer -menetelmä kuitenkin myös eroaa monilta osin BSiZerista. Seuraavissa luvuissa tarkastellaan tiheysfunktion estimointiin soveltuvan Bayes-SiZer -menetelmän toteutusta sekä vertaillaan sen antamia tuloksia SiZer-menetelmän tuloksiin.

4 Bayes-SiZer -menetelmä

4.1 Tiheysfunktion estimointi käyttäen Bayes-päättelyä

Oletetaan, että f on tuntematon tiheysfunktio, $X \sim f$ ja käytettävissä on satunnaisotos $\boldsymbol{X} = (X_1, \ldots, X_n)$, jolle pätee $X_1, \ldots, X_n \sim f$. Tiheysfunktion estimoinnissa pyritään muodostamaan tuntemattomalle tiheysfunktiolle $p_X(x)$ pisteessä x estimaatti $\hat{p}_X(x)$ hyödyntäen satunnaisotosta $\boldsymbol{X}$, jonka realisoitunutta arvoa merkitään

symbolilla $\boldsymbol{x} = (x_1, \ldots, x_n)$. Bayes-päättelyn merkinnöin notaation lyhentämiseksi tiheysfunktioon p liittyvän satunnaismuuttujan tai -vektorin merkintä jätetään kuitenkin jatkossa pois.

Olkoon $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ satunnainen parametrivektori. Tällöin Bayes-päättelyn kontekstissa parametrinen estimointi voidaan suorittaa muun muassa suurimman uskottavuuden menetelmällä, prediktiivisen jakauman avulla tai MAP-menetelmällä (engl. *maximum a posteriori probability*). Suurimman uskottavuuden menetelmässä etsitään ensin suurimman uskottavuuden estimaatti

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} p\left(\boldsymbol{x} \mid \boldsymbol{\psi}\right)$$

ja muodostetaan tämän avulla tiheysfunktiolle estimaatti $\hat{p}\left(x\right) = p\left(x \mid \hat{\boldsymbol{\psi}}\right)$. MAP-menetelmässä toimitaan samalla periaatteella, mutta siinä etsitään ensin suurimman posterioritodennäköisyyden estimaatti eli MAP-estimaatti

$$\boldsymbol{\psi}_{MAP} = \arg\max_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} p\left(\boldsymbol{\psi} \mid \boldsymbol{x}\right)$$

ja muodostetaan tämän avulla estimaatti $\hat{p}\left(x\right) = p\left(x \mid \boldsymbol{\psi}_{MAP}\right)$. Jos satunnaismuuttuja X tulkitaan uudeksi samasta jakaumasta generoiduksi havainnoksi ja oletetaan, että X on riippumaton aineistosta $\boldsymbol{X}$ ehdolla $\boldsymbol{\psi}$, niin voidaan muodostaa havainnnon X prediktiivinen tiheysfunktio

$$\hat{p}\left(x\right) = p\left(x \mid \boldsymbol{x}\right) = \int_{\boldsymbol{\Psi}} p\left(x \mid \boldsymbol{\psi}\right) p\left(\boldsymbol{\psi} \mid \boldsymbol{x}\right) \mathrm{d}\boldsymbol{\psi} \,. \tag{13}$$

Tällöin prediktiivinen tiheysfunktio hyödyntää koko parametriavaruutta $\boldsymbol{\Psi}$ painottamalla parametrisia malleja $p\left(x \mid \boldsymbol{\psi}\right)$ posterioritiheysfunktioilla $p\left(\boldsymbol{\psi} \mid \boldsymbol{x}\right)$. Jos satunnaisvektorin $\boldsymbol{\psi}$ posteriorijakaumasta on käytettävissä otos $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(m)}$, $m \geq 1$, voidaan estimaatin (13) lausekkeessa olevaa integraalia approksimoida Monte Carlo -menetelmällä, jolloin pätee

$$\hat{p}\left(x\right) \approx \frac{1}{m} \sum_{i=1}^{m} p\left(x \mid \boldsymbol{\psi}^{(i)}\right) .$$

Keskeisimmät tiheysfunktion estimointiin soveltuvat parametriset Bayes-menetelmät perustuvat joko normaalijakaumien sekoitteeseen (engl. *mixture of normals*) [19], wavelet-funktioihin [18] tai Logspline-menetelmään [11, 14, 15]. Yleisesti tarkasteltuna kaikki mainitut menetelmät mallintavat tuntematonta tiheysfunktiota muotoa

$g_j(\cdot\,;\boldsymbol{\alpha})$, $j = 1, \ldots, K$ olevien kantafunktioiden lineaarikombinaatiolla. Estimaatit ovat muotoa

$$p\,(x \,|\, \boldsymbol{\psi}) = p\,(x \,|\, \boldsymbol{\alpha}, \boldsymbol{\beta}, K) = C(\boldsymbol{\alpha}, \boldsymbol{\beta}, K) \sum_{j=1}^{K} \beta_j \, g_j(x\,;\boldsymbol{\alpha}), \quad x \in \mathbb{R}\,, \tag{14}$$

missä $C(\cdot)$ on normalisointitekijä ja $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ sekä $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ ovat kuhunkin menetelmään liittyviä parametrivektoreita. Mallista (14) nähdään, että silotteen sileyteen vaikuttavia tekijöitä ovat kantafunktioiden lukumäärä K ja parametrivektorit $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$.

Normaalijakaumien sekoitteessa kaavan (14) kantafunktiot $g_j(\cdot\,;\boldsymbol{\alpha})$ ovat muotoa $N(\mu_j, \sigma_j^2)$ olevien normaalijakaumien tiheysfunktioita. Wavelet-funktioiden mallissa kantafunktiot ovat puolestaan avaruuden $L^2(\mathbb{R})$ virittäviä wavelet-funktioita, ja Logspline-menetelmän mallissa kantafunktiot ovat luonnollisia kuutiollisia splinejä (engl. *natural cubic splines*). Näissä kaikissa menetelmissä pyritään muodostamaan tiheysfunktiolle kaavan (13) mukainen prediktiivinen estimaatti käyttäen apuna Markovin ketju Monte Carlo (MCMC)-algoritmeihin pohjautuvien otantamenetelmien tuottamia otoksia aineistolla $\boldsymbol{x}$ ehdollistetun satunnaisvektorin $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, K)$ posteriorijakaumasta. Näin ollen kaikki kolme menetelmää ovat Bayes-menetelmille ominaiseen tapaan hyvin laskentaintensiivisiä.

4.2 Logspline- ja Bayes-Logspline -menetelmä

Bayes-SiZer -menetelmän parametriseksi malliksi valitaan tässä työssä Logspline-menetelmän Bayes-päättelyä soveltava versio, jota tässä työssä kutsutaan Bayes-Logspline -menetelmäksi. Bayes-SiZer -menetelmän tarvitsemat silotteet voitaisiin toteuttaa myös muillakin edellä mainituilla menetelmillä, mutta Bayes-Logspline -menetelmä on idealtaan yksinkertainen eikä siinä tehdä oletuksia normaalijakautuneisuudesta. Seuraavaksi esitellään pääpiirteissään sekä Kooperbergin ja Stonen [14, 15] kehittämä Logspline-menetelmä että Hansenin ja Kooperbergin [11] Logspline-menetelmästä jalostama Bayes-Logspline -menetelmä. Splinifunktioita eli splinejä koskevan teorian osalta viitataan lähteeseen [20].

Logspline-menetelmässä oletetaan, että satunnaismuuttuja $X \sim f$ saa arvoja väliltä (L, U) ja rajoille sallitaan myös arvot $L = -\infty$ ja $U = +\infty$. Lisäksi oletetaan, että s on luonnollinen kuutiollinen splini solmupistein $t_1, \ldots, t_K$, missä $K \geq 4$. Tällöin splini s on korkeintaan ensimmäistä astetta oleva polynomi väleillä (L, t_1) ja

$[t_K, U)$. Merkitään luonnollisen kuutiollisen spliniavaruuden $\mathcal{NS}\,|_{(L,U)}$ kantafunktioita $1, B_1(\cdot\,;\boldsymbol{t}), \ldots, B_J(\cdot\,;\boldsymbol{t})$, missä $J = K-1$ ja $\boldsymbol{t} = (t_1, \ldots, t_K) \in \mathbb{R}^K$. Kooperberg ja Stone [14] konstruoivat kannan siten, että $B_1(\cdot\,;\boldsymbol{t})$ on lineaarinen negatiivisella kulmakertoimella välillä (L, t_1) ja vakio välillä $[t_K, U)$, $B_2(\cdot\,;\boldsymbol{t}), \ldots, B_{J-1}(\cdot\,;\boldsymbol{t})$ ovat vakioita väleillä (L, t_1) ja $[t_K, U)$ sekä $B_J(\cdot\,;\boldsymbol{t})$ on lineaarinen positiivisella kulmakertoimella välillä $[t_K, U)$ ja vakio välillä (L, t_1).

Olkoon $G \subset \mathcal{NS}\,|_{(L,U)}$ kantafunktioiden $B_1(\cdot\,;\boldsymbol{t}), \ldots, B_J(\cdot\,;\boldsymbol{t})$ virittämä aliavaruus, jolle pätee $\dim G = J$. Jos $g \in G$, niin se on muotoa

$$g(x\,;\boldsymbol{\beta},\boldsymbol{t}) = \beta_1 B_1(x\,;\boldsymbol{t}) + \cdots + \beta_J B_J(x\,;\boldsymbol{t})\,, \qquad x \in \mathbb{R}\,, \tag{15}$$

missä $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J) \in \mathbb{R}^J$. Tuntemattoman tiheysfunktion f logaritmin oletetaan nyt noudattavan mallia

$$\log f(x\,;\boldsymbol{\beta},\boldsymbol{t}) = C(\boldsymbol{\beta},\boldsymbol{t}) + g(x\,;\boldsymbol{\beta},\boldsymbol{t})\,, \qquad x \in \mathbb{R}\,. \tag{16}$$

Tällöin tiheysfunktiota koskevasta ehdosta (1) voidaan ratkaista termi $C(\boldsymbol{\beta},\boldsymbol{t})$, sillä

$$\begin{aligned}
1 &= \int_L^U \exp\big\{\log f(x\,;\boldsymbol{\beta},\boldsymbol{t})\big\}\,\mathrm{d}x = \int_L^U \exp\big\{C(\boldsymbol{\beta},\boldsymbol{t})\big\}\exp\big\{g(x\,;\boldsymbol{\beta},\boldsymbol{t})\big\}\,\mathrm{d}x \\
&\Leftrightarrow\ \exp\big\{C(\boldsymbol{\beta},\boldsymbol{t})\big\} = \left[\int_L^U \exp\big\{g(x\,;\boldsymbol{\beta},\boldsymbol{t})\big\}\,\mathrm{d}x\right]^{-1} \\
&\Leftrightarrow\ C(\boldsymbol{\beta},\boldsymbol{t}) = -\log\left[\int_L^U \exp\left\{\sum_{j=1}^J \beta_j B_j(x\,;\boldsymbol{t})\right\}\mathrm{d}x\right].
\end{aligned}$$

Jotta mallin (16) tuottama estimaatti olisi äärellinen, vaaditaan, että kerroinvektori $\boldsymbol{\beta}$ toteuttaa kiinteällä solmuvektorilla $\boldsymbol{t}$ ehdon $C(\boldsymbol{\beta},\boldsymbol{t}) < \infty$ tai ekvivalentisti molemmat ehdoista

(i) $L > -\infty$ tai $\beta_1 < 0$

(ii) $U < +\infty$ tai $\beta_J < 0\,.$

Olkoon $X_1, \ldots, X_n \sim f$ satunnaisotos. Logspline-menetelmässä hyödynnetään suurimman uskottavuuden menetelmää, jossa solmuvektorin $\boldsymbol{t}$ ollessa vakio maksimoidaan log-uskottavuusfunktiota

$$\ell(\boldsymbol{\beta},\boldsymbol{t}) = \sum_{i=1}^n \log f(X_i\,;\boldsymbol{\beta},\boldsymbol{t}) = \sum_{i=1}^n \sum_{j=1}^J \beta_j B_j(X_i\,;\boldsymbol{t}) + n\,C(\boldsymbol{\beta},\boldsymbol{t}) \tag{17}$$

parametrin $\boldsymbol{\beta} \in \mathcal{B}$ suhteen. Käytännössä maksimointi tapahtuu numeerisesti Newton-Raphson -menetelmän avulla [25]. Saatava tiheysfunktion estimaatti $\hat{f}_n(x\,;\,\hat{\boldsymbol{\beta}}, \boldsymbol{t})$ riippuu kuitenkin voimakkaasti käytetyistä solmupisteistä. Erityisesti solmujen lukumäärä K säätelee estimaatin silotteen määrää siten, että mitä pienempi K on sitä silotetumpi estimaatti saadaan. Solmujen lukumäärän ja sijaintien valitsemiseksi sovelletaan niin kutsuttua ahnetta etsintäalgoritmia (ks. [14]). Tätä ja muita samankaltaisia automaattisia mallinvalinta-algoritmeja on kuitenkin kritisoitu kaikkien mahdollisten mallien joukon liian suppeasta tutkimisesta.

Bayes-logspline -menetelmässä yritetään ratkaista solmupisteiden valintaan liittyvät ongelmat sisällyttämällä solmupisteiden valinta Bayes-päättelyn perustana olevaan todennäköisyysmalliin. Kokonaisuudessaan Bayes-Logspline -menetelmän todennäköisyysmalli muodostuu satunnaismuuttujasta X, aineistosta $\boldsymbol{X} = (X_1, \ldots, X_n)$, satunnaismuuttujasta K sekä satunnaisvektoreista $\boldsymbol{t} = (t_1, \ldots, t_K)$ ja $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$, missä $J = K - 1$.

Parametrivektorin $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{t}, K)$ priorijakauma voidaan kirjoittaa hierarkisessa muodossa

$$p\,(\boldsymbol{\psi}) = p\,(\boldsymbol{\beta} \,|\, \boldsymbol{t})\, p\,(\boldsymbol{t} \,|\, K)\, p\,(K)\,. \tag{18}$$

Kooperberg ja Hansen tutkivat artikkelissaan [11] synteettisten aineistojen avulla useita eri priorijakaumia. Näiden tarkastelujen tuloksiin perustuen tässä työssä on käytetty solmujen lukumäärälle K diskreettiä tasaisesti jakautunutta priorijakaumaa tiheysfunktiolla

$$p\,(K) = \begin{cases} \frac{1}{K_{\max} - K_{\min}}\,, & \text{kun } K_{\min} \leq K \leq K_{\max} \\ 0\,, & \text{muuten} \end{cases}\,, \tag{19}$$

missä $K_{\min}$ ja $K_{\max}$ valitaan sopivasti kuitenkin siten, että $4 \leq K_{\min} < K_{\max} < n$. Muita vaihtoehtoja parametrin K priorijakaumaksi ovat geometrinen jakauma sekä Poisson-jakauma, jota Denison ja hänen kollegansa käyttävät regressio-ongelman yhteydessä artikkelissa [4]. Prioritietämyksenä solmuvektorista $\boldsymbol{t}$ oletetaan, että kiinteällä solmupisteiden lukumäärällä K kaikki solmuvektorit ovat yhtä todennäköisiä ja solmupisteet saadaan kaikkien havaintopisteiden joukosta otannalla ilman takaisinpanoa. Voidaan siis kirjoittaa

$$p\,(\boldsymbol{t} \,|\, K) = \binom{n}{K}^{-1}. \tag{20}$$

Bayes-Logspline -menetelmän toteutuksessa solmuvektorin $\boldsymbol{t}$ avulla määritellään spliniavaruuden $\mathcal{NS}|_{(L,U)}$ kantafunktiot $B_1(\cdot\,;\boldsymbol{t}),\ldots,B_J(\cdot\,;\boldsymbol{t})$. Tämän jälkeen kerroinvektorin $\boldsymbol{\beta}$ priorijakaumaksi ehdolla $\boldsymbol{t}$ asetetaan osittain epäaito (engl. *partially improper*) multinormaalijakauma odotusarvolla $\mathbf{0}$ ja kovarianssimatriisilla $(\lambda\boldsymbol{A})^{-1}$, missä parametri $\lambda>0$ voidaan tulkita todennäköisyysmallin hyperparametriksi ja matriisi $\boldsymbol{A}$ rakentuu muotoa

$$A_{ij}=\int_L^U B_i''(x)B_j''(x)\,\mathrm{d}x\,,\qquad 1\le i,j\le J$$

olevista komponenteista (ks. [11]). Vaikka parametri λ voitaisiin sisällyttää todennäköisyysmalliin ja sille voitaisiin muodostaa hyperpriorijakauma, käyttävät Kooperberg ja Hansen kuitenkin kiinnitettyä arvoa. Heidän tarkasteluidensa perusteella tässäkin työssä päätettiin käyttää arvoa $\lambda=1/n$.

Bayes-Logspline -menetelmässä pyritään muodostamaan tuntemattomalle tiheysfunktiolle prediktiivinen estimaatti, mutta kaavassa (13) olevan posteriorijakauman

$$p\left(\boldsymbol{\psi}\mid\boldsymbol{x}\right)=\frac{p\left(\boldsymbol{x}\mid\boldsymbol{\psi}\right)p\left(\boldsymbol{\psi}\right)}{p\left(\boldsymbol{x}\right)}$$

tuntematon normalisointitekijä $p\left(\boldsymbol{x}\right)$ estää ratkaisemasta kaavan (13) integraalia analyyttisesti. Ongelman ratkaisemiseksi Bayes-Logspline -menetelmässä hyödynnetään Greenin [10] kehittämää Markovin ketju Monte Carlo -otantamenetelmän versiota nimeltä Reversible Jump MCMC (RJMCMC), jonka avulla voidaan poimia otos parametrivektorin $\boldsymbol{\psi}$ posteriorijakaumasta. RJMCMC-otantamenetelmän erikoisuutena on, että sen avulla voidaan poimia otos vaihtuvadimensioisen vektorin, kuten $\boldsymbol{\psi}$, jakaumasta. Perinteiset MCMC-menetelmät, kuten Gibbsin ja Metropoliksen–Hastingsin otantamenetelmä [3, 9], eivät näin ollen sovellu lainkaan käytettäviksi Bayes-Logspline -menetelmän yhteydessä. RJMCMC-otantamenetelmän käytön johdosta Bayes-Logspline -menetelmä on kuitenkin hyvin laskentaintensitiivinen. Laskennan vaatimuksia lisää se, ettei Logspline-menetelmän joustavuuden johdosta RJMCMC-otantamenetelmän toteutuksessa voida hyödyntää konjugaattijakaumia lainkaan.

4.3 Bayes-SiZer -menetelmän toteutus

SiZer-menetelmän periaatteiden mukaisesti ei myöskään Bayes-SiZer -menetelmässä pyritä tekemään päätelmiä todellisesta tiheysfunktiosta, vaan tarkastelut kohdistetaan tuntemattoman tiheysfunktion silotteisiin. Näiden avulla pyritään konstruoi-

maan menetelmän tulokseksi SiZer-värikarttoja tehokkaampia Bayes-SiZer -värikarttoja. Värikarttojen luonti kuitenkin edellyttää silotteiden riippumista vain yksiulotteisesta silotusparametrista ja Bayes-Logspline -menetelmässä, kuten muissakin edellä mainituissa parametrisissa Bayes-menetelmissä, silotteen sileyteen vaikuttavia tekijöitä ovat kaikki moniulotteisen parametrivektorin $\boldsymbol{\psi}$ komponentit. Siksi Bayes-SiZer -menetelmä toteutetaan kahdessa eri vaiheessa seuraavasti.

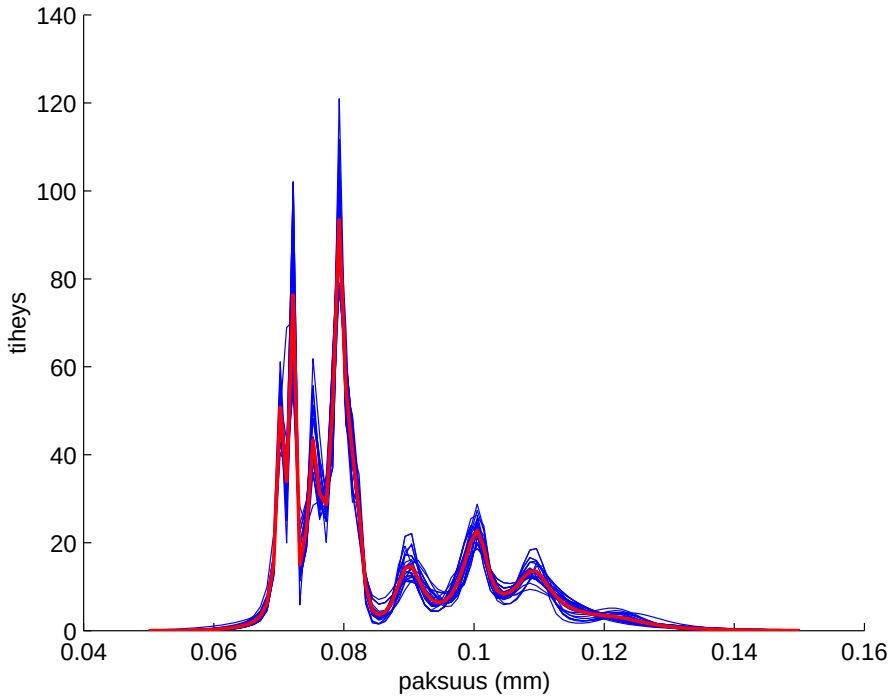
Bayes-SiZer -menetelmän ensimmäinen askel on poimia Bayes-Logspline -menetelmän ja tämän hyödyntämän RJMCMC-otantamenetelmän avulla tiheysfunktioiden perhe $\{p\,(x \mid \boldsymbol{\psi}^{(j)})\}_{j=1}^{m}$, missä m on suuri. Näiden tiheysfunktioiden tulkitaan edustavan parasta tietämystä tuntemattomasta tiheysfunktiosta, kun aineisto $\boldsymbol{X}$ tunnetaan. Toisessa askeleessa näistä tiheysfunktioista muodostetaan silotteita konvoloimalla ne Gaussin ytimen kanssa. Koska SiZer-menetelmässä päätelmät tehdään tarkasti ottaen silotteiden derivaatoista, konvoloidaan tiheysfunktiot Gaussin ytimen derivaatalla K_h'. Tällöin päättelyt kohdistetaan konvoloitujen silotteiden derivaattoihin ja SiZer-värikarttojen y-akselilla voidaan kuvata silotusparametrin h tai sen muunnoksen $\log_{10}(h)$ arvoja. Merkitään saatuja silotteita $\delta_h = p\,(\cdot \mid \boldsymbol{\psi}) * K_h'$. Silotteen δ_h konvoluutio evaluoidaan numeerisesti yli yksiulotteisen hilapisteikön $z_1 < \cdots < z_k$, ja samoja hilapisteitä käytetään myös Bayes-SiZer -menetelmän tuottamien värikarttojen evaluoinnissa.

Kuvassa 6 on eräitä Hidalgo-postimerkkiaineistosta Bayes-Logspline -menetelmällä tuotettuja estimaatteja $p\,(x \mid \boldsymbol{\psi}^{(j)})$ sekä näiden keskiarvokäyrä eli prediktiivinen tiheysfunktioestimaatti. Kuvassa 7 esitetään kuvan 6 estimaattien konvoluutiot Gaussin ytimen ja Gaussin ytimen derivaatan kanssa, kun $h = 0.0025$.

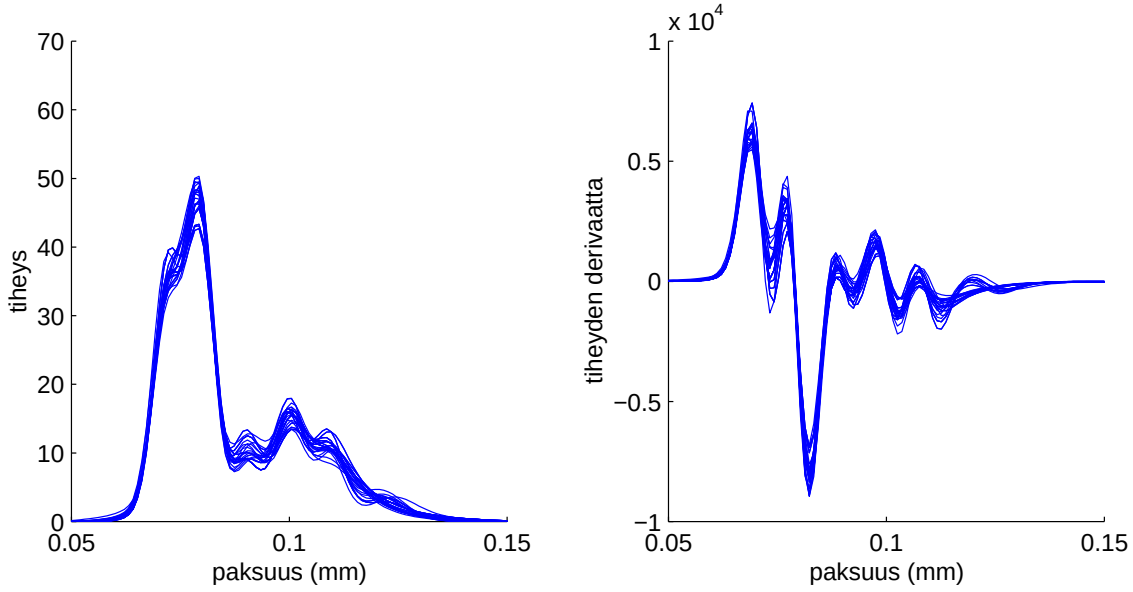
4.4 Bayes-SiZer -värikartat

Perinteisessä SiZer-menetelmässä piirteiden merkitsevyys ja jako värikartan eri väreihin pääteltiin silotteen derivaatan odotusarvon $\mathbb{E}\hat{f}_n'(\cdot\,;\,h)$ luottamusvälien avulla. Bayes-SiZer -menetelmässä käytettävä Bayes-päättely mahdollistaa kuitenkin todennäköisyyspäätelmien teon suoraan silotteen derivaatasta δ_h. Piirre tulkitaan siis merkitseväksi mikäli sen todennäköisyys on suurempi kuin $1-\alpha$, missä $\alpha \in (0, 1/2)$ on valittu vakio.

Bayes-SiZer -värikarttojen toteutus noudattaa samoja periaatteita kuin Erästön ja Holmströmin artikkelissa [7] kehittämä regressio-ongelmiin soveltuva, niin ikään



Kuva 6. Hidalgo-postimerkkiaineistosta Bayes-Logspline -menetelmällä tuotettuja posteriorikäyriä (sinisellä) sekä prediktiivinen tiheysfunktio (punaisella).



Kuva 7. Vasemmassa kuvassa Hidalgo-postimerkkiaineistosta Bayes-Logspline -menetelmällä tuotettujen posteriorikäyrien silotteet sekä oikeassa kuvassa silotteiden derivaatat silotusparametrin arvolla $h = 0.0025$.

Bayes-päättelyä hyödyntävä BSiZer. Bayes-SiZer -värikartoissa käytetyt värit ovat kuitenkin BSiZerista poiketen täysin samat kuin perinteisten SiZer-värikarttojen värit. Ainoa poikkeus SiZer-värikarttoihin on, ettei Bayes-SiZer -värikartoissa esiinny menetelmässä käytetyn Bayes-päättelyn johdosta harmaata väriä lainkaan.

Pisteittäisiin luottamusväleihin perustuvan SiZer-värikartan vastine saadaan jakamalla evaluointipisteiden indeksijoukko $\{1, \ldots, k\}$ kolmeen osajoukkoon

$$\begin{aligned} I^s &= \{i \,|\, \mathbb{P}(\delta_h(z_i) > 0) \geq 1 - \alpha\}\,, \\ I^p &= \{i \,|\, \mathbb{P}(\delta_h(z_i) < 0) \geq 1 - \alpha\}\,, \\ I^l &= \{1, \ldots, k\} \backslash (I^s \cup I^p)\,. \end{aligned} \tag{21}$$

Piste (z_i, h) värjätään siis värikartassa siniseksi, punaiseksi tai lilaksi sen mukaan kuuluuko i joukkoon I^s, I^p vai I^l.

Samanaikaisiin luottamusväleihin perustuvan SiZer-värikartan vastine saadaan jakamalla evaluointipisteiden indeksijoukko $\{1, \ldots, k\}$ kolmeen pistevieraaseen osajoukkoon J^s, J^p ja $J^l = \{1, \ldots, k\} \backslash (J^s \cup J^p)$ siten, että

$$\mathbb{P}(\delta_h(z_i) > 0, \text{ kun } i \in J^s \text{ ja } \delta_h(z_i) < 0, \text{ kun } i \in J^r) \geq 1 - \alpha\,. \tag{22}$$

Piste (z_i, h) värjätään siis värikartassa siniseksi, punaiseksi tai lilaksi sen mukaan kuuluuko i joukkoon J^s, J^p vai J^l. Todennäköisyysmitan $\mathbb{P}$ monotonisuudesta seuraa, että $J^s \subset I^s$ ja $J^p \subset I^p$, joten $J^l \supset I^l$. Erästö ja Holmström kuitenkin huomauttavat, että osajoukkoja J^s ja J^p ei voida valita yksikäsitteisesti [7]. He ehdottavat yhtenä ratkaisuna samanaikaisten Bayes-luottamusvälien (engl. *credible interval*) käyttöä.

Olkoon $\Delta > 0$ sellainen, että se toteuttaa ehdon

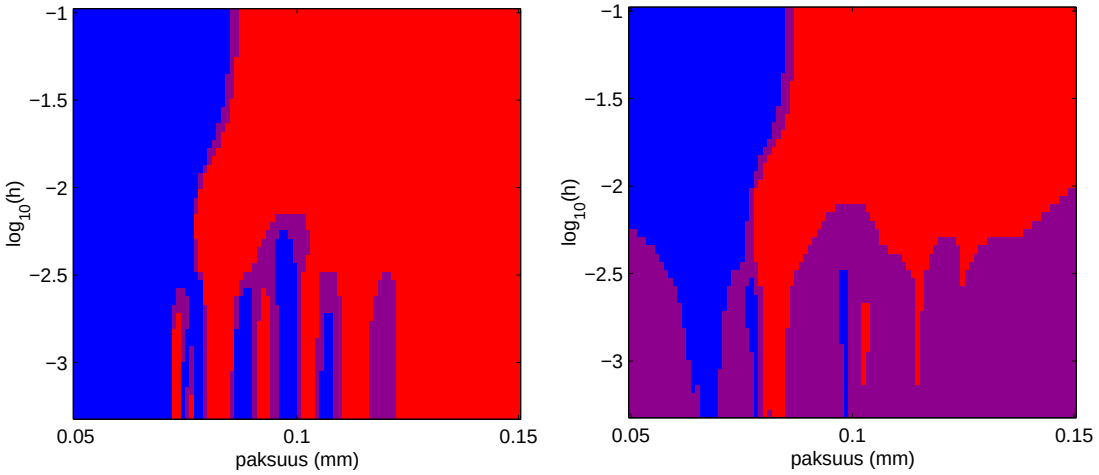
$$\mathbb{P}\left(\max_{i=1,\ldots,k} \left| \frac{\delta_h(z_i) - \mathbb{E}\delta_h(z_i)}{s(\delta_h(z_i))} \right| \leq \Delta \right) = 1 - \alpha\,, \tag{23}$$

missä $s(\delta_h(z_i))$ on termin $\delta_h(z_i)$ keskihajonta. Tällöin ehto (22) toteutuu, kun valitaan

$$\begin{aligned} J^s &= \{i \,|\, \mathbb{E}\delta_h(z_i) - \Delta\, s(\delta_h(z_i)) > 0\} \\ J^p &= \{i \,|\, \mathbb{E}\delta_h(z_i) + \Delta\, s(\delta_h(z_i)) < 0\}\,. \end{aligned} \tag{24}$$

Kaavojen (21) ja (23)-(24) todennäköisyyksien, odostusarvojen ja hajontojen evaluointi tapahtuu menetelmän numeerisessa toteutuksessa Bayes-Logspline -menetelmästä saatavan otoksen avulla.

Kuvassa 8 esitetään Hidalgo-postimerkkiaineiston avulla tuotetut alueisiin (21) ja (24) perustuvat Bayes-SiZer -värikartat. Kuten kuvan 5 SiZer-värikartoissa, myös nyt samanaikaiset luottamusvälit tuottavat konservatiivisemman värikartan kuin pisteittäiset luottamusvälit. Bayes-päättelyn ansiosta Bayes-SiZer -värikartat eivät siis sisällä harmaata väriä lainkaan mahdollistaen päätelmien teon myös Hidalgo-postimerkkiaineiston reuna-alueilla. Seuraavassa luvussa vertaillaan ja analysoidaan perinteisen SiZer-menetelmän ja Bayes-SiZer -menetelmän eroja tarkemmin synteettisten ja empiiristen aineistojen avulla.



Kuva 8. Hidalgo-postimerkkiaineistosta tuotetut Bayes-SiZer -värikartat. Vasemmanpuoleinen kartta on tuotettu käyttäen pisteittäisiä todennäköisyyksiä ja oikeanpuoleinen käyttäen samanaikaisia Bayes-luottamusvälejä. Molemmissa kartoissa on käytetty arvoa $\alpha = 0.05$.

5 Testausta

Tässä luvussa vertaillaan perinteisen SiZer-menetelmän ja Bayes-SiZer -menetelmän toimintaa sekä tuloksia yhden synteeettisen ja kahden empiirisen aineiston avulla. Koska menetelmien varsinaiset tulokset ovat värikarttoja, kohdistuvat tarkastelut lähinnä karttojen visuaalisiin eroihin. Luvussa 5.1 esitellään käytetyt aineistot, joiden valinnassa on pyritty huomioimaan, että hyviltä tilastollisilta menetelmiltä vaaditaan soveltuvuutta mahdollisimman erilaisiin aineistoihin. Varsinaisten tulosten tarkastelu esitetään kohdassa 5.2.

5.1 Aineistot

Bayes-SiZer -menetelmän testausta varten on valittu edellä esitetyn Hidalgo-postimerkkiaineiston lisäksi yksi synteettinen aineisto ja kaksi empiiristä aineistoa. Molemmat empiiriset aineistot ovat esiintyneet useasti tilastollisessa kirjallisuudessa eri menetelmien yhteydessä, mikä on osaltaan vaikuttanut aineistojen valintaan. Synteettisen aineiston perustana oleva tiheysfunktio on myöskin tunnettu, sillä se kuuluu Marronin ja Wandin [17] kehittämien 15 testitiheysfunktion joukkoon. Tiheysfunktio tunnetaan nimellä Marronin–Wandin tiheys #10 tai sen muotoa kuvaavalla englanninkielisellä nimellä *claw*. Marronin–Wandin tiheys #10 muodostuu kuuden normaalijakauman sekoitteesta, jossa normaalijakaumien tiheysfunktioille asetetaan odotusarvot, varianssit ja painot seuraavasti

$$\tfrac{1}{2}N(0,1) + \tfrac{1}{10}N(-1, \tfrac{1}{100}) + \tfrac{1}{10}N(-\tfrac{1}{2}, \tfrac{1}{100}) + \tfrac{1}{10}N(0, \tfrac{1}{100}) + \tfrac{1}{10}N(\tfrac{1}{2}, \tfrac{1}{100}) + \tfrac{1}{10}N(1, \tfrac{1}{100})\,.$$

Marronin–Wandin tiheys #10 on esitetty kuvan 9 vasemmassa ylänurkassa.

Ensimmäinen empiirisistä aineistoista käsittelee Yhdysvalloissa sijaitsevan Buffalon kaupungin lumitilannetta vuosina 1910-1972 (engl. *Buffalo snowfall data*). Aineisto koostuu 63 vuotuisesta lumen syvyyden (tuuma) yksiulotteisesta mittauksesta. Myös pienen kokonsa johdosta tämä aineisto soveltuu erinomaisesti SiZer- ja Bayes-SiZer -menetelmien vertailuun. Aineistoa on tutkittu esimerkiksi lähteessä [21], jossa se on myös esitetty kokonaisuudessaan. Jatkossa käytetään aineistosta nimeä Buffalo-aineisto.

Toinen empiirinen aineisto käsittelee Yhdysvalloissa Yellowstonen kansallispuistossa sijaitsevaa Old faithful -geysiriä (engl. *Old faithful geyser data*). Aineisto sisältää mittauksia geysirin 295 peräkkäisen purkauksen kestoista minuutteina mitattuna.

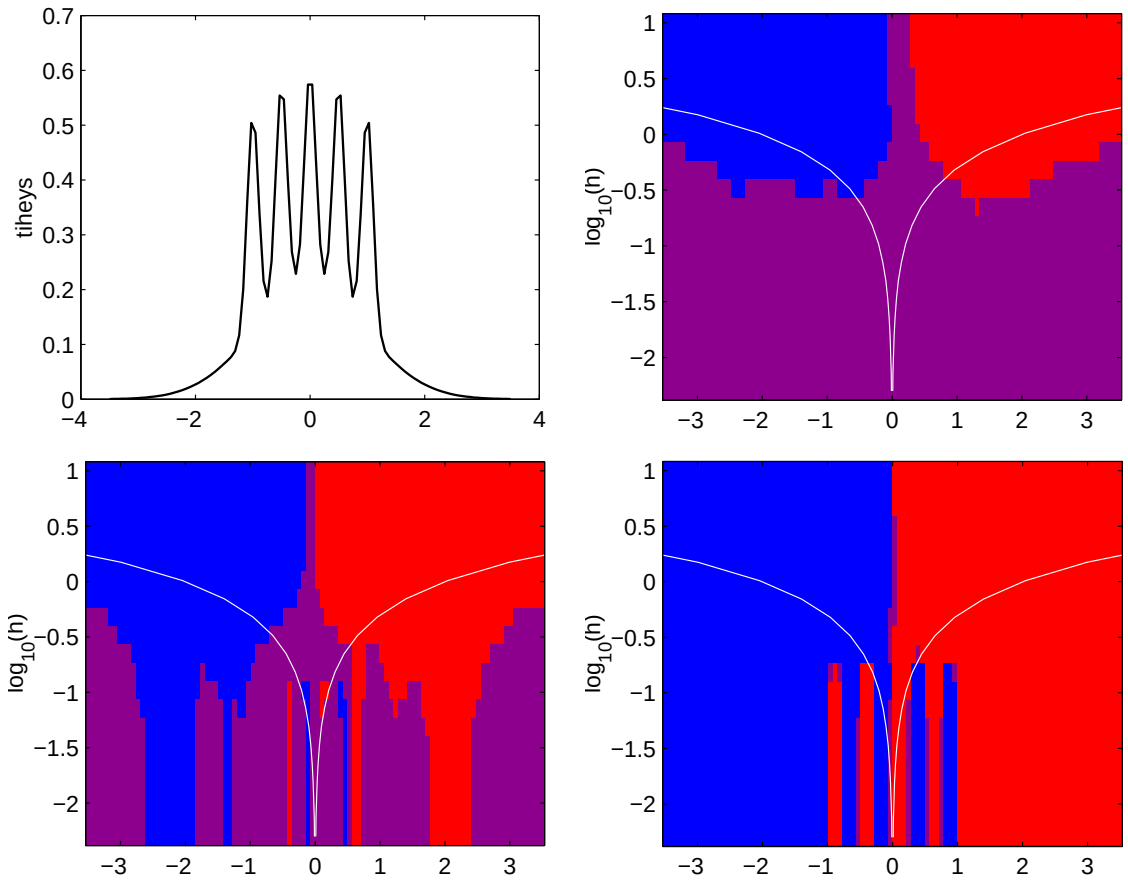
Vaikka nimi Old faithful viittaakin geysirin ikuiseen deterministiseen toimintaan, on sen purkauksissa kuitenkin havaittu muutoksia. Lisätietoa geysiristä saa Yellowstonen Internet-sivuilla `http://www.yellowstone.com`, missä geysirin toimintaa voi seurata myös reaaliaikaisesti. Aineisto on puolestaan ladattavissa Internet-osoitteesta `http://www.quantlet.org/mdbase/`. Jatkossa käytetään aineistosta nimeä geysir-aineisto.

5.2 Tulokset

Kuvassa 9 on esitetty Marronin-Wandin tiheys #10 sekä kolme Bayes-SiZer -värikarttaa, jotka on konstruoitu tiheyden #10 jakaumasta poimittujen erikokoisten otosten avulla. Oikeassa ylänurkassa oleva Bayes-SiZer -värikartta vastaa otoskokoa $n = 50$, vasemmassa alanurkassa oleva otoskokoa $n = 500$ ja oikeassa alanurkassa oleva otoskokoa $n = 5000$. Kaikki värikartat on tuotettu käyttäen luottamustason määrittelemiseksi arvoa $\alpha = 0.05$. Jokaista kuvan 9 Bayes-SiZer -värikarttaa varten poimittiin Bayes-Logspline -menetelmällä 1000 suuruinen otos parametrivektorin $\boldsymbol{\psi}$ posteriorijakaumasta. Konvergenssin todettiin tapahtuneen 200 iteraation jälkeen, joten lopulliseksi posterioriotoksen kooksi jäi $m = 800$. Värikarttoihin on myös lisätty päätelmien teon helpottamiseksi ja silottamisen määrän havainnollistamiseksi kaksi valkoista käyrää, joiden etäisyys toisistaan on $4h$. Tämä perustuu siihen, että silotusparametrilla h Gaussin ytimen K_h voidaan tulkita olevan oleellisilta osin positiivinen välillä $[-2h, 2h]$ ja nolla välin ulkopuolella. Näiden käyrien esittäminen kuuluu myös perinteiseen SiZer-menetelmään.

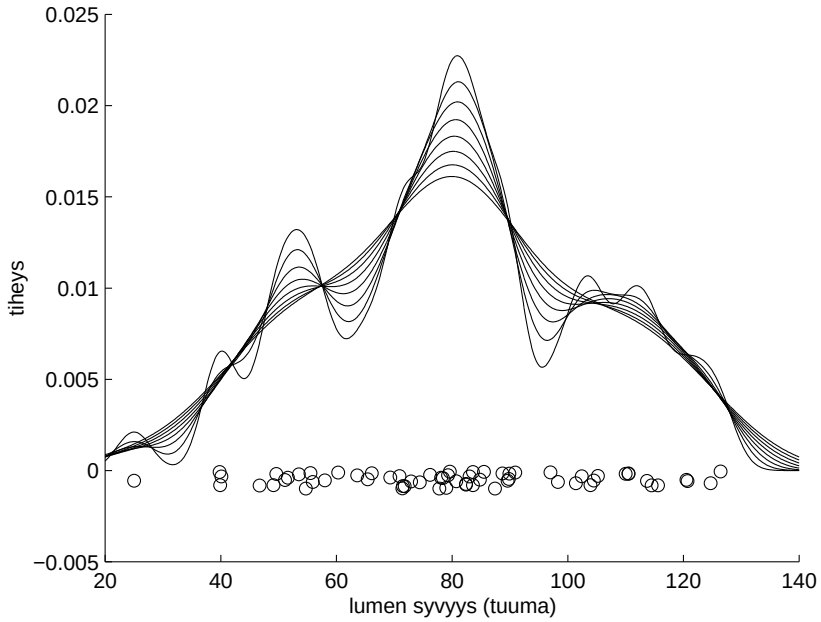
Tarkastelemalla kuvan 9 Bayes-SiZer -värikarttoja nähdään selvästi otoskoon n vaikutus merkitsevien piirteiden löytämiseen. Otoskoon kasvaessa lilan värin osuus pienenee värikartoisssa ja merkitsevien moodien lukumäärä kasvaa yhdestä todelliseen lukumäärään viisi. Otoskoon ollessa $n = 50$ nähdään, että vähäinen silottaminen ei tuo merkitseviä piirteitä esiin lainkaan. Vasta otoskoolla $n = 500$ saadaan varmuutta reunimmaisten moodien sekä keskimmäisen moodin olemassaolosta. Otoskoko $n = 5000$ puolestaan riittää mainiosti kaikkien moodien löytämiseen.

Kuvassa 10 on esitetty kahdeksan Buffalo-aineistosta ydinestimaattorin avulla tuotettua silotetta. Ytimenä on käytetty Gaussin ydintä ja silotusparametrille h on annettu arvoja tasaisesti väliltä $[3, 10]$. Silotteen sileydestä riippuen, voidaan havaita moodien lukumäärän vaihtelevan yhden ja seitsemän välillä. SiZer-menetelmien avulla pyritään selvittämään kuinka moni moodeista todella on merkitsevä.

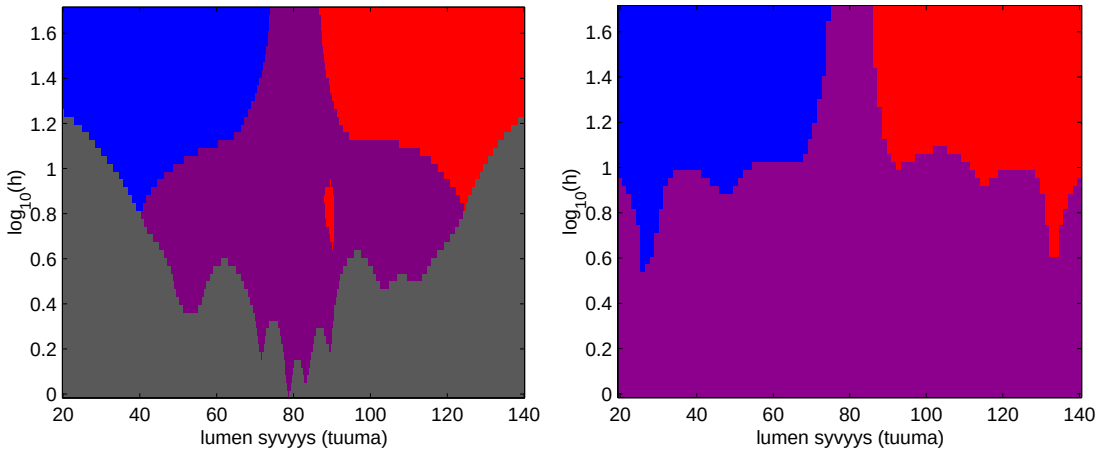


Kuva 9. Vasemmassa yläkulmassa on Marronin tiheysfunktio numero 10. Oikealla ylhäällä olevan Bayes-SiZer -värikartan tekoon on käytetty otoskokoa $n = 50$, vasemmalla alhaalla otoskoko on $n = 500$ ja oikealla alhaalla $n = 5000$. Kaikki värikartat on toteutettu käyttäen samanaikaisia Bayes-luottamusvälejä arvolla $\alpha = 0.05$. Värikarttoihin on lisätty myös Gaussin ytimen efektiivisen kantajan ilmoittava (valkoinen) käyrä.

Kuvan 11 Buffalo-aineiston avulla tuotetuista värikartoista vasemmanpuoleinen on tuotettu SiZer-menetelmällä ja oikeanpuoleinen Bayes-SiZer -menetelmällä. Molemmat värikartat on tuotettu käyttäen luottamustason määrittelemiseksi arvoa $\alpha = 0.05$. Vasemmanpuoleisesta värikartan nojalla SiZer-menetelmä tulkitsee ainoastaan kohdassa 80 tuumaa olevan moodin merkitseväksi. Tämä päätelmä voidaan siis tehdä kaikilla silotteilla, joille pätee $\log_{10}(h) \geq 0.8$. Tätä pienemmille silotusparametrin arvoille SiZer tunnistaa ainoastaan merkitsevän laskun kohdassa 90 tuumaa. Harmaan värin osuus värikartassa on kuitenkin kohtalaisen suuri. Tämä johtuu osaltaan aineiston pienuudesta, sillä perinteinen SiZer-menetelmä ei tällöin pysty tekemään päätelmiä silotteiden derivaatoista. Violetin värin alue kertoo hyvin, miten aineiston havainnot harvenevat kohdan 80 tuumaa ympäriltä kohti aineiston reuna-alueita. Violetin värin alueella SiZer ei havaitse merkitsevää laskua eikä nousua silotteiden derivaatoissa.



Kuva 10. Buffalo-aineistosta tuotettuja ydinestimaatteja eri silotusparametrin arvoilla. Ytimenä on käytetty Gaussin ydintä. Buffalo-aineiston havainnot on merkitty kuvaan y-akselin arvon 0 alapuolelle symbolilla ”o”. Havaintoja on täristetty y-akselin suuntaisesti havaintojen toisistaan erottamisen helpottamiseksi.



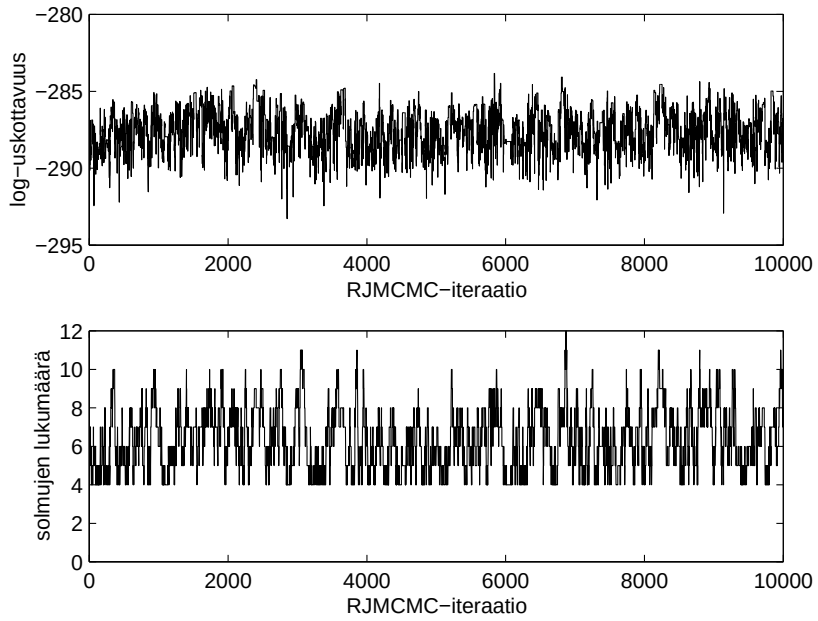
Kuva 11. Buffalo-aineistosta tuotetut värikartat. Vasemmanpuoleinen kartta on tuotettu SiZer-menetelmällä käyttäen samanaikaisia luottamusvälejä ja oikeanpuoleinen kartta on tuotettu Bayes-SiZer -menetelmällä käyttäen samanaikaisia Bayes-luottamusvälejä. Molemmissa kartoissa on käytetty arvolla $\alpha = 0.05$ saatavaa luottamustasoa.

Buffalo-aineiston tapauksessa SiZer-menetelmän heikkous on, että silotteilla, joille pätee $\log_{10}(h) < 0.8$, ei harmaan alueen johdosta voida päätellä olevan ainuttakaan 95 % luottamustasolla merkitsevää moodia. Kuitenkin esimerkiksi kuvan 10 vähiten sileällä silotteella ($h = 3$ eli $\log_{10}(h) \approx 0.5$) havaitaan selkeästi olevan useita moodeja.

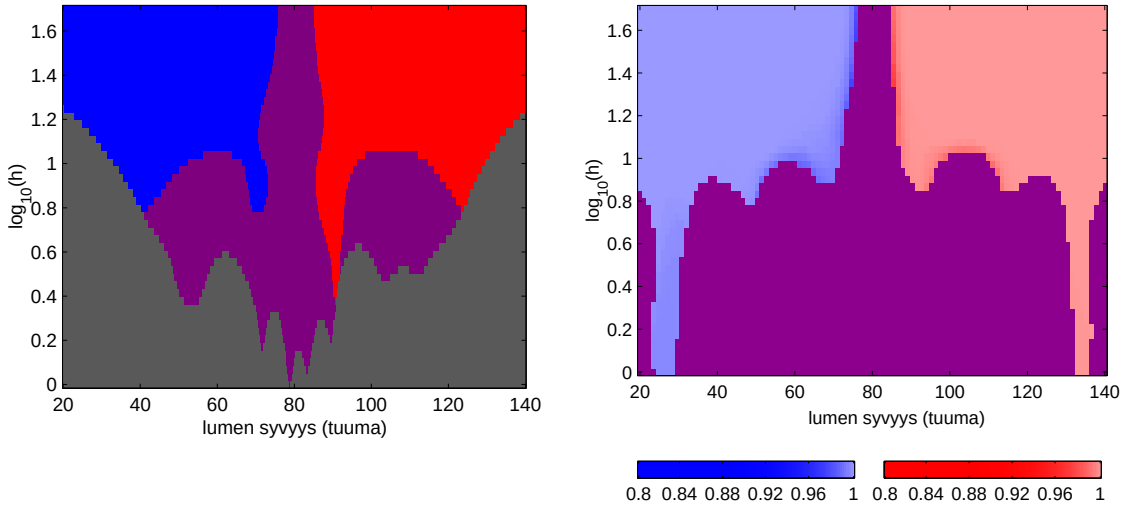
Kuvan 11 oikeanpuoleisen värikartan toteutusta varten poimittiin Bayes-Logspline -menetelmällä 10000 suuruinen otos parametrivektorin $\boldsymbol{\psi}$ posteriorijakaumasta. RJMCMC-algoritmin konvergenssia tarkasteltiin visuaalisesti hyödyntäen kuvan 12 kaltaisia tilastoja, joissa on kuvattu mallien log-uskottavuuksien ja solmujen lukumäärien käyttäytymistä suhteessa algoritmin iteraatioihin. Näiden tarkastelujen perusteella ketjun pääteltiin saavuttaneen tasapainojakaumansa jo 500 iteraation jälkeen, joten lopullisen otoksen kooksi saatiin $m = 9500$. Tämän otoksen avulla muodostetusta värikartasta nähdään, että myös Bayes-SiZer -menetelmän mukaan kohdan 80 tuumaa moodi on lumen syvyyden tiheysfunktion ainoa merkitsevä moodi. Bayes-SiZer -värikartta ei kuitenkaan kärsi harmaan alueen tietämättömyydestä, vaan päätelmät on mahdollista tehdä y-akselin koko arvoalueella. Kuitenkin myös Bayes-SiZer -värikartta tulkitsee pienimpien silotusparametrien arvoilla tiheysfunktion olevan vailla moodia. Kuvan 13 värikartoista kuitenkin nähdään kuinka molemmat menetelmät pystyvät parempaan päättelyyn myös pienimmillä silotusparametrin arvoilla, kun vaadittua luottamustasoa pienennetään eli parametrin α arvoa suurennetaan. Kuvan 13 oikeanpuoleisessa Bayes-SiZer -värikartassa on sovellettu artikkelin [7] ideaa ja lisätty sinisiin ja punaisiin alueisiin väriskaalat ilmaisemaan vastaavien todennäköisyyksien suuruuksia.

Kuvassa 14 esitetään viisi geysir-aineistosta tuotettua ydinestimaattia, joissa ytimenä on käytetty Gaussin ydintä. Kuvan 14 jokaiseen silotteeseen on myös merkitty käytetyn silotusparametrin h arvo. Huomattavaa on, että vaikka silotusparametrin arvoa $h = 0.1$ vastaavassa silotteessa näyttäisi olevan kaksi melkein yhtäsuurta moodia, niin silotusparametrin h suurentuessa vasemmanpuoleinen moodi jää aina pienenmmäksi kuin oikeanpuoleinen moodi. Tämä havainto herättää kysymyksiä vasemmanpuoleisen moodin olemasssaolosta ja merkitsevyydestä.

Kuvan 15 geysir-aineiston avulla tuotetuista värikartoista vasemmanpuoleinen on tuotettu SiZer-menetelmällä ja oikean puoleinen Bayes-SiZer -menetelmällä. Molemmat värikartat on tuotettu käyttäen luottamustason määrittelemiseksi arvoa



Kuva 12. RJMCMC-algoritmin konvergenssitilastoja Buffalo-aineistolla.

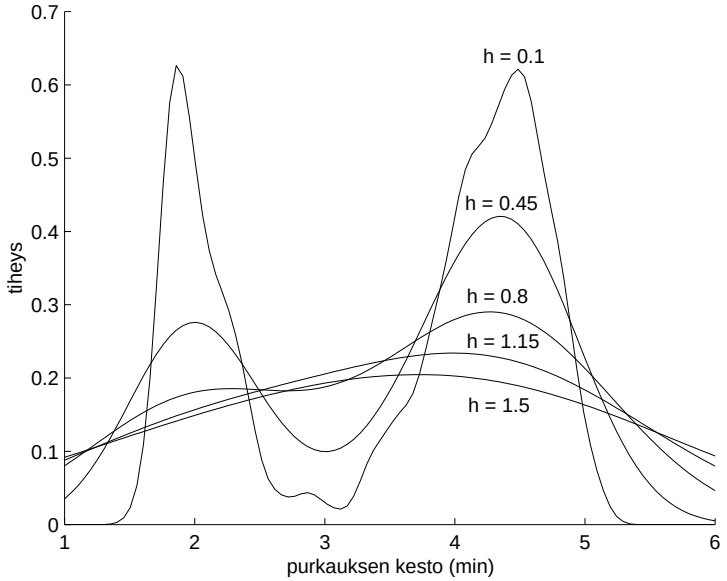


Kuva 13. Buffalo-aineistosta tuotetut värikartat. Vasemmanpuoleinen kartta on tuotettu SiZer-menetelmällä käyttäen samanaikaisia luottamusvälejä, kun taas oikeanpuoleinen kartta on tuotettu Bayes-SiZer -menetelmällä käyttäen samanaikaisia Bayes-luottamusvälejä. Molemmissa kartoissa on käytetty arvolla $\alpha = 0.2$ saatavaa luottamustasoa.

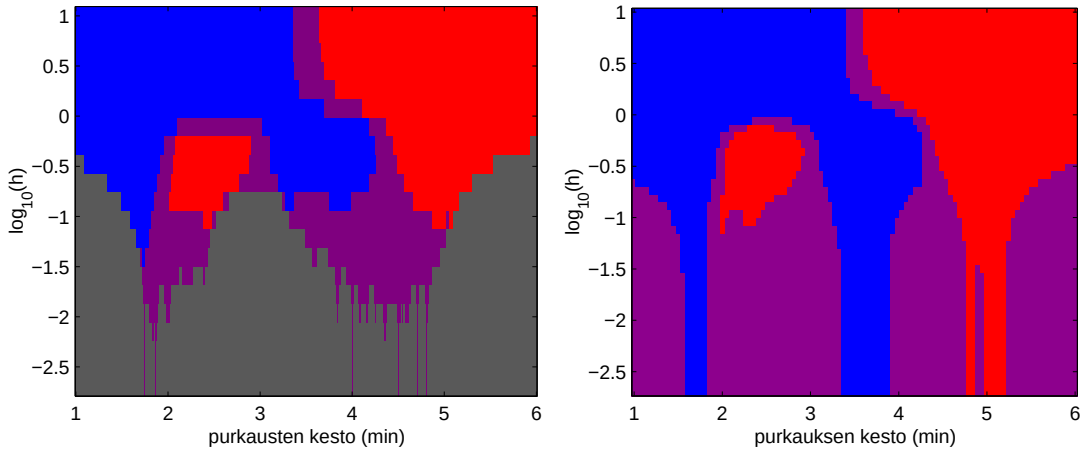
$\alpha = 0.05$. Sileimmillä silotteilla SiZer-menetelmä tunnistaa geysirin purkauksen kestolle vain yhden moodin $3\frac{1}{2}$ minuutin kohdalla. Jos silotusta vähennetään, niin SiZer-

menetelmä löytää kaksi moodia kohdista 2 ja $4\frac{1}{2}$ minuuttia. Kohdassa 2 minuuttia olevan moodin merkitsevyys pienillä silotusparametrin arvoilla jää kuitenkin SiZer-värikartassa epäselväksi, sillä alueessa $\log 10(h) < -1$ päätelmien teko on harmaan alueen johdosta epävarmaa.

Kuten Buffalo-aineiston tapauksessa, Bayes-SiZer -värikarttaa varten poimittiin Bayes-Logspline -menetelmällä 10000 suuruinen otos parametrivektorin $\boldsymbol{\psi}$ posteriorijakaumasta. Visuaalisten tarkastelujen perusteella ketjun pääteltiin saavuttaneen tasapainojakaumansa 2000 iteraation jälkeen. Lopullisen otoksen kooksi jäi siten $m = 8000$. Tämän otoksen avulla muodostettu värikartta yhtyy perinteisen SiZer-menetelmän tekemiin päätelmiin geysir-aineiston piirteistä alueessa $\log 10(h) \geq -1$. Bayes-SiZer -menetelmällä päättelyt voidaan kuitenkin ulottaa luotettavasti myös alueelle $\log 10(h) < -1$. Tällä alueella Bayes-SiZer -menetelmä ei kuitenkaan enää tulkitse kohdassa 2 minuuttia olevaa moodia riittävän merkitseväksi. Menetelmä tunnistaa edelleen silotteiden derivaatoissa merkitsevän nousun kohdassa 1.8 minuuttia mutta ei sen jälkeistä laskua. Kohdasa $4\frac{1}{2}$ minuuttia olevan moodin merkitsevyydestä ei sen sijaan ole epäselvyyttä.



Kuva 14. Geysir-aineistosta tuotettuja ydinestimaatteja eri silotusparametrin arvoilla. Ytimenä on käytetty Gaussin ydintä.



Kuva 15. Geysir-aineistosta tuotetut värikartat. Vasemmanpuoleinen kartta on tuotettu SiZer-menetelmällä käyttäen samanaikaisia luottamusvälejä, kun taas oikeanpuoleinen kartta on tuotettu Bayes-SiZer -menetelmällä käyttäen samanaikaisia Bayes-luottamusvälejä. Molemmissa kartoissa on käytetty arvolla $\alpha = 0.05$ saatavaa luottamustasoa.

6 Johtopäätökset

Edellä esitettyjen esimerkkien tapauksissa Bayes-SiZer -menetelmällä on saavutettu parannuksia SiZer-menetelmän suoritukseen. Hyödyntämällä Bayes-päättelyä on voitu välttää SiZer-menetelmässä esiintyvät normaaliapproksimaatiot ja löytää näin aineistosta piirteitä alueilla, joissa se ei SiZer-menetelmällä ole mahdollista. Lisäksi mahdollisuus tehdä päätelmiä käyttäen hyvinkin pieniä silotusparametrin arvoja mahdollistaa päätelmien teon silotteiden sijasta suoraan tuntemattomasta tiheysfunktiosta itsestään.

Tiheysfunktion estimointiin tarkoitetuissa Bayes-menetelmissä ei juurikaan voida hyödyntää konjugaattijakaumia (vrt. [7]). Tämän johdosta posterioriotoksen poimintaan käytetyt MCMC-otantamenetelmät ovat laskennallisesti hyvin raskaita. Bayes-SiZer -menetelmässä laskentaintensiivisyys näkyy Bayes-Logspline -menetelmän kautta tehden Bayes-SiZer -menetelmän paljon perinteistä SiZer-menetelmää hitaammaksi. Näin ollen Bayes-SiZer -menetelmä ei sovellu perinteisen SiZer-menetelmän tavoin nopeaksi työkaluksi data-analyysiin. Laskennallista nopeutta voidaan saavuttaa Bayes-Logspline -menetelmään hyödyntämällä tässä työssä käytetyn ohjelmakoodia tulkkaavan MATLAB-ohjelmiston sijaan jotakin matalan tason käännettävää ohjelmointikieltä kuten C, C++ tai Fortran. Lisäksi posterioriotoksen poimin-

ta Bayes-Logspline -menetelmän sijaan jollakin muulla Bayes-menetelmällä, kuten artikkelin [19] normaalijakaumien sekoitteeseen perustuvalla menetelmällä, saattaisi parantaa Bayes-SiZer -menetelmän laskennallista tehokkuutta. Tämän mahdollisuuden selvittäminen jätetään jatkotutkimusten tehtäväksi.

Vaikka MCMC-otantamenetelmien toteutusta voitaisiinkin nopeuttaa, eivät ne sovellu käytettäviksi mustien laatikoiden tavoin rutiineina, joiden suoritusta ei tarvitsisi valvoa. Esimerkiksi ketjujen konvergenssiin ja hyperparametrien valintaan liittyviin ongelmiin on aina kiinnitettävä huomiota jokaisen aineiston kohdalla erikseen. Tiheysfunktion estimoinnissa Bayes-SiZer -menetelmää voi kuitenkin suositella käytettäväksi esimerkiksi tilanteissa, joissa halutaan tarkempaa tietoa niihin kysymyksiin, joihin perinteinen SiZer-menetelmä ei anna vastausta. Mikäli muutenkin data-analyysissä käytetään Bayes-menetelmin saatuja tiheysfunktioestimaatteja, onnistuu Bayes-SiZer -menetelmän käyttö tällöin nopeasti jo kerättyjen otosten avulla.

Viitteet

[1] P. Chaudhuri ja J. S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

[2] P. Chaudhuri ja J. S. Marron. Scale space view of curve estimation. *Annals of Statistics*, 28:408–428, 2000.

[3] D. G. T. Denison, C. C. Holmes, B. K. Mallick, ja A. F. M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. J. Wiley & Sons, Chichester, 2002.

[4] D. G. T. Denison, B. K. Mallick, ja A. F. M. Smith. Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society Series B*, 60(2):333–350, 1998.

[5] L. Devroye. *A Course in Density Estimation*. Birkhäuser, Boston, 1987.

[6] G. Elfving ja P. Tuominen. *Todennäköisyyslaskenta II*, ss. 177–179. Limes ry, Helsinki, toinen laitos, 1990.

[7] P. Erästö ja L. Holmström. Bayesian multiscale smoothing for making inferences about features in scatter plots. Lähetetty julkaistavaksi, 2003.

[8] J. Fan ja I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London, 1996.

[9] A. Gelman, J. B. Carlin, H. S. Stern, ja D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, toinen laitos, 2003.

[10] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[11] M. H. Hansen ja C. Kooperberg. Spline adaptation in extended linear models. *Statistical Science*, 17(1):2–51, 2002.

[12] L. Holmström ja P. Erästö. Making inferences about past environmental change using smoothing in multiple time scales. *Computational Statistics and Data Analysis*, 41(2):289–309, 2002.

[13] A. J. Izenman ja C. J. Sommer. Philatelic mixtures of multimodal densities. *Journal of the American Statistical Association*, 83(404):941–953, 1988.

[14] C. Kooperberg ja C. J. Stone. A study of logspline density estimation. *Computational Statistics and Data Analysis*, 12:327–348, 1991.

[15] C. Kooperberg ja C. J. Stone. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1:301–328, 1992.

[16] T. Lindeberg. Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(3):234–254, 1990.

[17] J. S. Marron ja M. P. Wand. Exact mean integrated squared error. *Annals of Statistics*, 20(2):712–736, 1992.

[18] P. Müller ja B. Vidakovic. Bayesian inference with wavelets: Density estimation. *Journal of Computational and Graphical Statistics*, 7(4):456–468, 1998.

[19] K. Roeder ja L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439), 1997.

[20] L. L. Schumaker. *Spline Functions: Basic Theory*, ss. 309–316. J. Wiley & Sons, New York, 1981.

[21] D. W. Scott. *Multivariate Density Estimation. Theory, Practice, and Visualization.* J. Wiley & Sons, Inc., New York, 1992.

[22] A.C.-C. Shih, H.-Y.M. Liao, ja Chun-Shien Lu. A new iterated two-band diffusion equation: theory and its application. *IEEE Transactions on Image Processing*, 12(4):446–476, 2003.

[23] B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society Series B*, 43(1):97–99, 1981.

[24] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York, 1986.

[25] M. A. Tanner. *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Springer-Verlag, New York, toinen laitos, 1993.